

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
6 December 2001 (06.12.2001)

PCT

(10) International Publication Number
WO 01/92859 A1

(51) International Patent Classification⁷: **G01N 21/64**,
A61B 5/00

(DK). **HARRIT, Niels** [DK/DK]; Enebærvej 3, DK-2400
Copenhagen NV (DK). **BRO-JØRGENSEN, Rasmus**
[DK/DK]; Strandvejen 64, DK-4600 Køge (DK).

(21) International Application Number: PCT/DK01/00383

(22) International Filing Date: 1 June 2001 (01.06.2001)

(74) Agent: **HØIBERG APS**; St. Kongensgade 59B, DK-1264
Copenhagen K (DK).

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
PA 2000 00863 2 June 2000 (02.06.2000) DK

(71) Applicant (for all designated States except US): **MEDI-
COMETRICS APS** [DK/DK]; Høstvej 7, DK-2920 Char-
lottenlund (DK).

(81) Designated States (*national*): AF, AG, AI., AM, AT, AU,
AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU,
CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH,
GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC,
LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW,
MX, MZ, NO, NZ, PI., PT, RO, RU, SD, SE, SG, SI, SK,
SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA,
ZW.

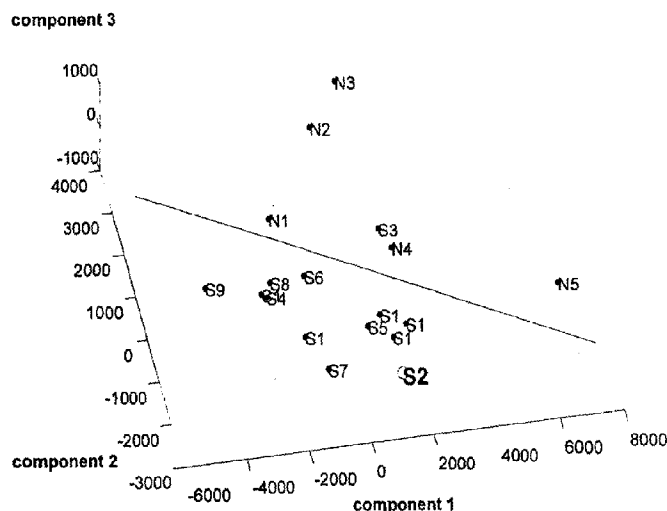
(72) Inventors; and

(75) Inventors/Applicants (for US only): **NØRGAARD,
Lars** [DK/DK]; Strandvejen 34, 2.tv., DK-2900 Hellerup
(DK). **ALBRECHTSEN, Morten** [DK/DK]; Høstvej
7, DK-2920 Charlottenlund (DK). **OLSEN, Ole, Inge-
mann** [DK/DK]; Skovvej 77 D, DK-2920 Charlottenlund

(84) Designated States (*regional*): ARIPO patent (GH, GM,
KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian
patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European
patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE,
IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF,
CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: METHOD AND SYSTEM FOR CLASSIFYING A BIOLOGICAL SAMPLE



(57) Abstract: The present invention relates to a method of training a classification system for characterising a biological sample, a diagnostic classification system, as well as a method of characterising a condition in an animal or a human being by using parameters obtained from the sample. The invention relates to classification based on physical parameters obtained from luminescence spectroscopy on light emitted from the sample. The data obtained from a spectrofluorimetric analysis can be considered a finger-print of the sample. Each sample gives rise to a unique spectrofluorometric set of physical parameters. By analysing the fluorescence data, it is possible to classify samples into two or more classes based on the fluorescence spectra, such as classifying with respect to presence/absence of a specific disease, group of diseases or risk of later attaining a specific disease or a body condition, or concentration of a specific compound or medicine.



WO 01/92859 A1

WO 01/92859 A1



Published:

— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Method and system for classifying a biological sample

The present invention relates to a method of training a classification system for characterising a biological sample, a diagnostic classification system, as well as a
5 method of characterising a condition in an animal or a human being by using parameters obtained from the sample.

Background

10 A need for a fast and reliable primary diagnostic tool providing information indicative of a disease or a group of diseases has existed for years.

In US 4,755,684 (Leiner et al.) a method for tumor diagnosis by means of serum tests is disclosed. The method includes excitation of the serum by an excitation radiation at least of a wavelength between 250 nm and 300 nm, and its fluorescence
15 intensity is measured at predetermined emission wavelengths. From deviations of these measuring values, a conclusion may be drawn with respect to the presence of a neoplastic disease. Measurements at one or two excitation wavelengths are suggested. Up to three emission wavelengths are determined for each excitation wavelength and an intensity value is determined. Since very little information from the
20 fluorescence spectroscopy is used the diagnosis is very rough and insecure. Only about 60 % are diagnosed correctly and the diagnosis is limited to a yes or no.

In WO 96/30746 and WO 98/24369 fluorescence spectra are used to screen tissue samples in situ, wherein the tissue suspected to be dysplastic tissue is directly subjected to fluorescence spectroscopy. The methods are used to distinguish between
25 dysplastic cervical tissue and normal cervical tissue. In O'Brien K.M. et al "Development and evaluation of spectral classification algorithms for fluorescence guided laser angioplasty", IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, vol. 36, No. 4, April 1989, pages 424-4430, fluorescence spectroscopy is used to distinguish normal arterial tissue from atherosclerotic tissue. None of these methods allows a specific diagnosis to be made based on analysis of spectra from tissue or
30 body fluids not directly related to the diseased tissue.

In US 5,734,587 a method of analyzing sample liquids by generating infrared spectra of dried samples and evaluating using a multivariate evaluation procedure is disclosed. In the evaluation procedure the samples are assigned to classes. The evaluation procedure is trained with samples of known classes to adjust the parameters of the evaluation procedures, such that samples of unknown classification can be assigned to known classes. The samples analysed are clinically relevant liquid samples, that have to be dried before generating the infrared spectra of the samples due to the nature of infrared spectra.

Most organic compounds absorb light in the visible or ultraviolet part of the electromagnetic spectrum. Many molecules emit the absorbed excitation energy in the form of fluorescence. A fluorescence spectrum is obtained by transmitting light to the sample (excitation light) and determining the spectral distribution of the light emitted from the sample. In the case where only one fluorescent compound is present in a weakly absorbing solution, the spectral profile of the fluorescence will be invariant with respect to the excitation wavelength. Only the intensity of the fluorescence will vary with the wavelength of the excitation light in accordance with the absorption spectrum.

If more than one fluorescent compound is present in the solution the relation between excitation and emission intensities will rapidly increase to a very high level of complexity. The individual compounds will absorb differently for each excitation wavelength, the intensity and distribution of the fluorescence will vary with excitation wavelength, and reabsorption of emitted photons might occur.

When a series of fluorescence spectra using different excitation wavelengths are recorded, the spectra collected represents an emission-excitation-matrix (EEM), which can be displayed as a 3-dimensional landscape (Figure 1). The EEM is specific for the specific mixture of compounds and the conditions under which it is measured.

Summary

It has been an object of the present invention to provide a method capable of classifying samples with unknown properties in a system not requiring any drying, en-

richment, separation or concentration of the sample before determining the class, to which the sample belongs.

5 This has been possible by subjecting the sample to fluorescence spectroscopy or a variant thereof, whereby liquid as well as solid samples may be classified.

Thus, in a first aspect the present invention relates to a method of training a classification system for characterising a biological sample with respect to at least one condition, comprising

10

a) obtaining a biological sample from an animal, including a human, wherein said biological sample is selected from body fluids and/or tissue, wherein the tissue sample is not associated with said condition(s),

15

b) obtaining characterisation information related to each biological sample,

c) exposing the sample to excitation light within a predetermined range of wavelength,

20

d) determining physical parameter(s) of light emitted from the sample,

e) repeating step a) to d) until the physical parameters of all training samples have been determined,

25

f) optionally performing a data handling of the obtained physical parameters obtaining data variables,

g) optionally performing a multivariate data analysis of the data variables obtaining model parameters describing the variation of the data variables,

30

h) classifying the biological samples into at least two different classes correlated to the characterisation information, obtaining a trained classification system.

In a preferred embodiment the method comprises the steps of:

- 5 a) obtaining a biological sample from an animal, including a human, wherein said biological sample is selected from body fluids and/or tissue, wherein the tissue sample is not associated with said condition(s),
- b) obtaining characterisation information related to each biological sample,
- 10 c) exposing the sample to excitation light within a predetermined range of wavelength,
- d) determining physical parameter(s) of light emitted from the sample,
- 15 e) repeating step a) to d) until the physical parameters of all training samples have been determined,
- f) performing a data handling of the obtained physical parameters obtaining data variables,
- 20 g) optionally performing a multivariate data analysis of the data variables obtaining model parameters describing the variation of the data variables,
- 25 h) classifying the biological samples into at least two different classes correlated to the characterisation information, obtaining a trained classification system.

In another preferred embodiment the method comprises the steps of:

- 30 a) obtaining a biological sample from an animal, including a human, wherein said biological sample is selected from body fluids and/or tissue, wherein the tissue sample is not associated with said condition(s),
- b) obtaining characterisation information related to each biological sample,
- 35

- 5
- c) exposing the sample to excitation light within a predetermined range of wavelength,
 - d) determining physical parameter(s) of light emitted from the sample,
 - e) repeating step a) to c) until the physical parameters of all training samples have been determined,
 - f) performing a data handling of the obtained physical parameters obtaining data variables,
 - 10 g) performing a multivariate data analysis of the data variables obtaining model parameters describing the variation of the data variables,
 - 15 h) classifying the biological samples into at least two different classes correlated to the characterisation information, obtaining a trained classification system.

20 In another aspect the present invention relates to a classification system for characterising a biological sample, said system comprising:

- a) a sample domain for comprising a biological sample,
- b) light means for exposing the sample to excitation light in the sample domain,
- 25 c) a detecting means recording the physical parameter(s) of light emitted from the sample,
- d) optionally computing means for performing data handling of the physical parameters, obtaining data variables,
- 30 e) optionally processing means for providing model parameters from data variables of the sample,
- 35 f) at least one storage means for storing physical parameters and/or data variables and/or model parameters of the biological sample,

- 5
- g) at least one storage means for storing physical parameters and/or data variables and/or model parameters and characterisation information of a trained classification system,
 - h) means for correlating physical parameters and/or data variables and/or model parameters from the sample with physical parameters and/or data variables and/or model parameters of the trained system, and
 - 10 i) means for displaying the characterisation class(es) of a sample.

In a preferred embodiment the system comprises:

- 15
- a) a sample domain for comprising a biological sample,
 - b) light means for exposing the sample to excitation light in the sample domain,
 - c) a detecting means recording the physical parameter(s) of light emitted from the sample,
 - 20 d) computing means for performing data handling of the physical parameters, obtaining data variables,
 - e) optionally processing means for providing model parameters from data variables of the sample,
 - 25 f) at least one storage means for storing physical parameters and/or data variables and/or model parameters of the biological sample,
 - 30 g) at least one storage means for storing physical parameters and/or data variables and/or model parameters and characterisation information of a trained classification system,

- h) means for correlating physical parameters and/or data variables and/or model parameters from the sample with physical parameters and/or data variables and/or model parameters of the trained system, and
- 5 i) means for displaying the characterisation class(es) of a sample.

In another preferred embodiment the system comprises:

- 10 a) a sample domain for comprising a biological sample,
- b) light means for exposing the sample to excitation light in the sample domain,
- c) a detecting means recording the physical parameter(s) of light emitted from the sample,
- 15 d) computing means for performing data handling of the physical parameters, obtaining data variables,
- e) processing means for providing model parameters from data variables of the sample,
- 20 f) at least one storage means for storing physical parameters and/or data variables and/or model parameters of the biological sample,
- g) at least one storage means for storing physical parameters and/or data variables and/or model parameters and characterisation information of a trained classification system,
- 25 h) means for correlating physical parameters and/or data variables and/or model parameters from the sample with physical parameters and/or data variables and/or model parameters of the trained system, and
- 30 i) means for displaying the characterisation class(es) of a sample.

In yet another aspect the invention relates to a method for characterising a biological sample of an animal, including a human, comprising

- 5 a) obtaining a biological sample from the animal or human,
- b) exposing the sample to excitation light,
- c) determining the physical parameter(s) of light emitted from the sample,
- 10 d) optionally performing a data handling of the obtained physical parameters obtaining data variables,
- e) storing the physical parameters and/or data variables and/or model parameters,
- 15 f) optionally providing model parameters from data variables of the sample,
- g) obtaining physical parameters and/or data variables and/or model parameters from a trained classification system,
- 20 h) correlating physical parameters and/or data variables and/or model parameters from the sample with physical parameters and/or data variables and/or model parameters of the trained system, and
- 25 i) displaying characterisation class(es) of the sample.

In yet another aspect the invention relates to a method for characterising a biological sample of an animal, including a human, comprising

- 30 a) obtaining a biological sample from the animal or human,
- b) exposing the sample to excitation light,
- c) determining the physical parameter(s) of light emitted from the sample,

- d) performing a data handling of the obtained physical parameters obtaining data variables,
- 5 e) storing the physical parameters and/or data variables and/or model parameters,
- f) optionally providing model parameters from data variables of the sample,
- 10 g) obtaining physical parameters and/or data variables and/or model parameters from a trained classification system,
- h) correlating physical parameters and/or data variables and/or model parameters from the sample with physical parameters and/or data variables and/or model parameters of the trained system, and
- 15 i) displaying characterisation class(es) of the sample.

In yet another aspect the invention relates to a method for characterising a biological sample of an animal, including a human, comprising

- 20 a) obtaining a biological sample from the animal or human,
- b) exposing the sample to excitation light,
- 25 c) determining the physical parameter(s) of light emitted from the sample,
- d) performing a data handling of the obtained physical parameters obtaining data variables,
- 30 e) storing the physical parameters and/or data variables and/or model parameters,
- f) providing model parameters from data variables of the sample,

- g) obtaining physical parameters and/or data variables and/or model parameters from a trained classification system,
- h) correlating physical parameters and/or data variables and/or model parameters from the sample with physical parameters and/or data variables and/or model parameters of the trained system, and
- i) displaying characterisation class(es) of the sample.

Thus, the comparison of the sample and the classification information in the trained classification system can be carried out on different levels of data, namely by comparing either the physical parameters and/or the data variables and/or the model parameters. It is likewise conceivable that two of the levels of data or all three levels can be used in the comparison of the biological sample to the classification information in the trained classification system.

According to the first aspect of the invention, namely the method of training a classification system, step b) which relates to obtaining classification information related to each biological sample can be carried out at any point in time as long as the information is available for the last step (step h) of the training method.

According to a preferred embodiment of the three aspects of the invention, the model parameters are latent variables being weighted averages of the data variables.

The method is preferably carried out in a classification system trained according to the present invention.

Drawings

Figure 1. Fluorescence landscape of typical urine sample. Intensity is given as a function of excitation and emission wavelength.

Figure 2. Three-dimensional score plot of latent variable (component) one versus two versus three. The 18 samples are labelled according to smoker/non-smoker (S/N) and person (number).

Figure 3. Typical fluorescence excitation-emission landscape from a sample from a fasting person

- 5 Figure 4. A scatter plot of score one, two and five from a PCA model of 23 fasting and non-fasting persons.

Figure 5. Score scatter plots from a PCA model of data from persons with benign tumors. The plots show score 1 versus 2, 1 versus 3 and 2 versus 3.

- 10 Figure 6. A plot of the raw fluorescence data used in the analysis. The 28 spectra from each sample are arranged successively on an arbitrary wavelength scale.

- 15 Figure 7. Influence-plot of samples in two latent dimensions (bold) and in three latent dimensions (ordinary font). The behaviour of sample five going from poor description (high residual) to high impact on the model (high leverage) is indicative of an outlying behaviour. This sample is not visible as an outlier in the plot of the raw data (Figure 6).

- 20 Figure 8. Score-plot showing the samples from the cardiac patients investigation in terms of the first principal component versus the second principal component.

- 25 Figure 9. Unfolded variable averaged fluorescence spectra for the 8 samples. The first emission top corresponds to excitation at 230 nm and the last emission top corresponds to excitation at 500 nm.

Figure 10A. PC1 vs. PC2 score plot from a PCA on auto scaled data. Figure 10B. PC1 score from a PCA on auto scaled data.

- 30 Figure 11A PC1 vs. PC2 score plot from a PCA on mean centered data. Figure 11B PC1 score from a PCA on mean centered data.

- 35 Figure 12A Predicted vs. measured for log(concentration) with all bacteria samples (i.e. without control sample). Figure 12B Predicted vs. measured for log(concentration) without the control sample and the sample containing 10^8 cells.

Figure 13. Upper part: Front face fluorescence spectrum of an undiluted blood plasma sample. Lower part: Front face fluorescence spectrum of the same sample diluted 1:5000. Notice the different intensity scales.

5

Figure 14A. Excitation 230 nm as a function of the dilution. 2 is diluted 1:5000, 3 is 1:3000, 4 is 1:2000, 5 is 1:700, 6 is 1:500, 7 is 1:200, 8 is 1:100, 9 is 1:50, 10 is 1:25, 11 is 1:10, 12 is 1:5, 13 is 1:2, 14 is undiluted sample. Figure 14B Excitation 250 nm as a function of the dilution. Same dilutions as in Figure 14A. (Measured in front face mode).

10

Figure 15A Excitation 310 nm as a function of the dilution. Same dilutions as in Figure 14A. Excitation 360 nm as a function of the dilution. Same dilutions as in Figure 14A. (Measured in front face mode).

15

Figure 16A. PC1 vs. PC2 score plot from a PCA. Figure 16B. PC3 vs. PC4 score plot from a PCA

20

Figure 17. Upper part: Transmission fluorescence spectrum of an undiluted blood plasma sample. Lower part: Transmission fluorescence spectrum of the same sample diluted 1:5000. Notice the different intensity scales.

Figure 18A. Excitation 230 nm as a function of the dilution. 2 is diluted 1:5000, 3 is 1:3000, 4 is 1:2000, 5 is 1:1000, 6 is 1:700, 7 is 1:500, 8 is 1:200, 9 is 1:100, 10 is 1:50, 11 is 1:25, 12 is 1:10, 13 is 1:5, 14 is 1:2, 15 is undiluted sample. Figure 18B. Excitation 250 nm as a function of the dilution. Same dilutions as in Figure 18A. (Measured in transmission mode).

Figure 19A. Excitation 310 nm as a function of the dilution. Same dilutions as in Figure 18A. Figure 19B. Excitation 360 nm as a function of the dilution. Same dilutions as in Figure 18A. (Measured in transmission mode).

Figure 20A. PC1 vs. PC2 score plot from a PCA. Figure 20B. PC3 vs. PC4 score plot from a PCA.

Figure 21A. Predicted vs. measured for front face for 1:25 to 1:5000. Figure 21B Predicted vs. measured for transmission for 1:25 to 1:5000.

Figure 22A. PC1 vs. PC2 score plot from a PCA on transmission samples. A, B, C, and D are different buffers with pH values of approx. 8.5-9.0, 7.0-7.5, 5.0-6.5, and 0.1 M HCl, respectively. Numbers like 1:2 indicate the dilution factor. Figure 22B PC1 vs. PC2 score plot from a PCA on front face samples.

Figure 23 PC1 vs. PC2 from a PCA on all samples. F is front face and T is transmission

Figure 24. Unfolded fluorescence spectra at four different pH values for transmission mode samples diluted 1:10. The difference observed for 0.1 M HCl spectrum seen in insert B is not observed for the other dilutions.

Detailed description of the invention

The invention relates to classification based on physical parameters obtained from the luminescence spectroscopy on light emitted from the sample. For practical reasons most of the discussion in this description relates to fluorescence spectroscopy. However as described below other physical parameters may be used in the classification. Thus, throughout the description the term fluorescence is used as an

equivalent of any luminescence type and is to be interpreted as such, unless disappropriate in specific embodiments.

5 Fluorescence spectroscopy is an extremely sensitive tool. The data obtained from a spectrofluorimetric analysis can be considered a finger-print of the sample. Each sample gives rise to a unique spectrofluorometric set of physical parameters, however, as is described by the present invention. When analysing the fluorescence data, it has become possible to classify samples into two or more classes based on the fluorescence spectra, if there is any systematic difference between the samples.

10 The difference between the samples will mostly not relate to a single component or a few components of the sample, but rather to a combination of a wide variety of components. This combination exhibits a pattern so complex that it is detectable by multivariate analysis only.

15 Thus, according to the evaluation of the fluorescence parameters it is possible to obtain more information about a biological sample, than it is when evaluating the various chemical components in the sample individually, i.e. it is possible to obtain inter-component information. Furthermore, there is no need to know the exact composition of components in the sample, as it is the fluorescence finger-print rather

20 than the components of the sample that is detected. If so desired, in a specific application, it may be possible to give a chemical characterisation of the information used by the classification system. It may even in certain situations be possible to do so directly from the mathematical parameters derived from the physical parameters.

25 For one sample normally more than several thousand data variables are obtained, and the amount of data increases by the number of samples used but the number of data variables is constant for each sample. In prior art it was common practise to discard most of the spectrofluorimetric information and use but a few selective or semi-selective physical parameters but the present invention makes use of all avail-

30 able information.

By the present invention it has become possible to obtain information regarding an animal or a human being by subjecting a biological sample from said animal (human being) to a fluorescence analysis. Examples of the information provided by the present invention may be any information regarding health condition, such as informa-

35

tion regarding presence/absence of a specific disease, group of diseases or risk of later attaining a specific disease or a body condition, or concentration of a specific compound or medicine.

5 In a first aspect the invention relates to a method of training a classification system for characterising a biological sample. It is the purpose of the training that a classification system is obtained, said system holding enough information to be used for characterising an un-classified and unknown biological sample into one of the classes of the classification system. By the term unknown is meant a sample for
10 which no characterisation information is known.

It is also the purpose of the training of the system that this training incorporates a validation that substantiates how well classification can be performed on specific samples in the future as well as improving the validation specificity and sensitivity
15 over time.

Samples

The biological sample may be any sample suitable for fluorescence analysis. The
20 sample may be fluid or solid, as is appropriate. It is an object of the present invention to acquire the necessary information from the sample using as few pre-treatments as possible, preferably without any pre-treatments as such.

Accordingly, in a most preferred embodiment the sample is transferred directly from
25 the animal or human being to be subjected to fluorescence analysis, in order to obtain data relating to fresh, un-treated samples. In case it is not possible to use the biological sample directly, it may be stored, for example by freezing the sample.

A characteristic of the biological sample is that it is preferably not directly related to
30 the specific conditions, in that the spectroscopy is preferably not conducted on the tissue suspected to express the disease, whereby it is often possible to diagnose a condition or a disease in a easy manner, since the biological sample to be examined may be easily established, such as a urine sample or a plasma sample.

Fluid samples may be any fluid samples obtainable from animals or human beings, i.e. body fluids, such as biological samples selected from blood, plasma, serum, saliva, urine, cerebrospinal fluid, tears, nasal secrete, semen, bile, lymph, milk, sweat and/or faeces.

5

In a preferred embodiment the fluids are easily available fluids, such as urine samples, milk, blood and/or serum and/or plasma samples. Most preferred are urine, milk or saliva samples or any other samples that are obtainable without any invasive technique.

10

The fluid sample is subjected to fluorescence analysis without drying, and preferably without any other changes in concentration, such as separation and enrichment. The fluid sample may be arranged in a sample compartment being closed or open before exposing the sample to excitation light.

15

It is however also possible with the present invention to use tissue samples, such as solid tissue samples directly. Examples of tissue samples include hair and nails. The tissue sample may be any sample, such as a biopsy of tissue, that is subjected to fluorescence spectroscopy. In the present invention the tissue is not directly related to the specific condition(s), thus the term "the tissue sample is not associated with said condition(s)" means that for example when classifying with respect to cancer the tissue sample does not represent the possibly cancerous tissue, but tissue from another part of the individual.

20

25 The biopsy may be from any tissue, such as from muscle, cutis, subcutis, kidney, brain, and liver.

The solid samples may be classified on the solid form, but it may often be necessary to provide a liquid form of the tissue before subjecting the sample to fluorescence spectroscopy. The liquid form of the tissue may be obtained by dissolving the tissue or mechanically destroying the tissue, such as blending the tissue, to obtain a suitable liquid suspension of the tissue.

30

Furthermore, it is possible to use a sample positioned in situ or non-invasive, i.e. not removed from its normal environment. The sample does not need to be physi-

35

cally removed from its place in the animal or human being. The invention also encompasses the possibility of in situ analysis of samples. This can be done easily with samples like skin, hair, and nails, but it is likewise possible to conduct the excitation and fluorescence light beams by means of light guides to and from the liquid or tissue samples within the body. This may be accomplished by conducting the measurements transdermally. The light guides may thus be introduced into the body via body openings, such as the mouth, nose, ears, rectum, vagina, or urethra, or the light guides may be introduced through the blood vessels or inserted directly into tissue. In this way various fluids may be measured in situ, as well as some of the solid tissue samples.

In a preferred embodiment the biological sample is selected from body fluids, hair and nails, more preferred from body fluids.

Excitation light

The physical parameters may in principle be obtained for a wide variety of excitation light wavelengths. The wavelengths are preferably selected to be within the range of from 100 nm to 1000 nm, such as from 100 to 800 nm, more preferably within the range of from 200 nm to 800 nm, such as from 200 nm to 600 nm.

Normally several wavelengths are used, such as from 2 to 10,000, 4 to 10,000, 2 to 1000, 4 to 1000, 2 to 100 wavelengths, such as from 4 to 100 wavelengths, for instance 2-30, such as from 4 to 30 wavelengths, such as 2-10, such as from 4 to 10 wavelengths, for instance 2-6 wavelengths in order to describe an excitation-emission matrix optimally. Sets of wavelength may be chosen so that each wavelength differs from the other by at least 0.1 nm, such as at least 0.5 nm, for instance at least 1 nm, such as at least 5 nm, for instance at least 10 nm, such as at least 50 nm, for instance at least 100 nm, such as at least 150 nm, for instance at least 250 nm, such as at least 500 nm, for instance at least 600 nm, such as at least 700 nm, and at most 750 nm.

Multiwavelength excitation may be established either sequentially by varying the setting of a monochromator or other dispersing or filtering device in front of a continuous lightsource like a xenon lamp. Alternatively, the sample may be exposed to the full spectrum of a continuous light source equipped with a polychromator which

disperses the light spatially. Thus, different zones of the sample are exposed to exciting light of different wavelengths. Furthermore, an array of single wavelength light sources light e.g. lasers or light guide bundles may be used either in the sequential mode or in the spatially separated mode.

5

Accordingly, at least 2 excitation light wavelengths are selected such as at least 4, at least 6, at least 8, at least 10, or more. The excitation light of each wavelength may be used simultaneously or sequentially. In a preferred embodiment 4 wavelengths are selected, such as excitation light having a wavelength of 230 nm, 240 nm, 290 nm, and 340 nm. Each sample is then subjected to excitation light of each wavelength. In another preferred embodiment 6 wavelengths are selected.

10

The predetermined excitation light wavelength(s) is provided by use of light sources as is known to a person skilled in fluorescence spectroscopy.

15

Emission

The determination of the various physical parameters is done by equipment known to the person skilled in the art.

20

In fluorescence spectroscopy emission light intensities at different wavelengths are recorded for each excitation light wavelength. Preferably the emission light is sampled with 0.5 nm intervals or 1 nm intervals. Thereby a matrix of excitation-emission data is obtainable for each sample. Normally the spectral distribution of light emitted from the sample is ranging from 200 nm to 800 nm.

25

The emitted light is detected by any suitable detector, such as a one-dimensional detector, for example a photomultiplier. Alternatively, a scanning camera, a diode array, a CCD or a CMOS, all in principle being viewed as a two-dimensional array of several thousand or more detectors. The intensity of the light is detected on each detector thus permitting the whole spectrum or the whole EEM to be obtained in a single electronic measurement.

30

The emitted light from the samples may be focused onto the detectors by means of conventional focusing systems, as well as passing through diaphragms and mirrors.

Physical parameters

5

Most frequently the physical parameter to be determined in order to perform a data analysis is the intensity as a function of excitation wavelength and/or the emission wavelength. However, any other information contained in photoluminescence may be obtained from the sample such as fluorescence lifetime, phosphorescence intensity, phosphorescence lifetime, polarisation, polarisation lifetime, anisotropy, anisotropy lifetime, phase-resolved emission, circularly polarised fluorescence, fluorescence-detected circular dichroism, and any time dependence of the two last mentioned parameters.

10

Fluorescence intensity is easily measured at room temperature, and may therefore be chosen for many of the samples. Furthermore, a great number of organic natural products are known to be fluorescent. Phosphorescence may, however, also be performed at room temperature.

15

Luminescence lifetime in general, as well as phosphorescence lifetime are defined as the time required for the emission intensity to drop to $1/e$ of its initial value.

20

When using phase resolved fluorescence spectroscopy it is possible to suppress Raman and scattered light, leading to very good results for multicomponent systems.

25

In luminescence polarisation measurements, conventional spectra are obtained by scanning excitation spectra and measuring intensity parallel and perpendicular to the polarisation of exciting light. The polarisation may be calculated as the ratio of the difference of the two measurements to the sum of the two measurements. The anisotropy parameter is obtained by multiplying the perpendicular intensity by two in the denominator sum of this ratio.

30

Processing

The detectors are preferably coupled to a computer for further processing of the data. The physical parameters measured or determined by the detector are processed to a form suitable for the further mathematical calculations. This is done by allocating data variables to each physical parameter determined, thus obtaining data variables related to the physical parameters,

The physical parameters determined are often subjected to a data analysis through the data variables, such as a one-way matrix of spectral information, a two-way matrix of spectral information, a three-way matrix of spectral information, a four-way matrix of spectral information or, a five-way or higher-order matrices of spectral information.

Characterisation information

To obtain information relating to a specific condition in the animal or human being it is of importance that the data relating to the spectra obtained are correlated to characterisation information regarding the same biological sample. The information regarding the biological sample is preferably obtained substantially simultaneously with the biological sample, however, for characterisation data not varying essentially it is sufficient to obtain the data after having obtained the physical parameters.

The characterisation information relates to the classes intended to be generated through the training period.

The characterisation information is for instance relating to the presence or absence of a physical condition, such as a specific disease, or information regarding smoking, drinking, abuse of drugs, nutritive condition, etc. Furthermore, the characterisation information may include information such as sex, race, age or the like that is relevant for the classification. The characterisation information may also be information regarding responsiveness to a treatment as well as information regarding side effects of a treatment. The characterisation information may give information of both qualitative and quantitative information.

Also predictive information regarding an individual's risk of acquiring a condition or disease may be obtained by the present invention. The training of the system may be conducted by subjecting a kohorte of individuals to successive sample analysis and classify the samples into groups of individuals acquiring the disease and groups staying healthy during the period of sampling.

The characterisation information must be correlated to the spectral information obtained from the sample, in order to obtain the trained system ready for testing unknown samples.

10

Validity

Each sample is subjected to fluorescence spectroscopy before the data analysis is performed in the training of the classification system. The sample may be one sample from each animal or human being, or several samples from the same individual, each sample obtained at a different time interval or from different fluids or from different instruments.

15

Depending on the classes to be identified when training the classification system it is of importance to train the system with a sufficient number of samples. The determination of the sufficient number of samples is primarily determined by the similarity of the fingerprints of different classes. Indirectly, this can often be related to the number of expected latent variables, wherein the latent variables are weighted averages of the data variables. It is preferred that the ratio of number of training samples to the expected number of latent variables is at least 5:1, preferably at least 10:1. More preferred the ratio is 50:1, and even more preferred 100:1. The more training samples, the more reliable a system. Training is a continual improvement of the system and any sample is also a training sample being weighted decreasingly, however, over time.

20

25

30

The samples being classified in each class are preferably a representative group of samples to allow the most reliable classification, wherein representative is meant to mean exhibiting all variations influencing said classification. These variables can for example be age, sex, medication, existing disease, and race to match the population for which the classification system is designed.

35

Mathematics

5 A central aspect of the invention is the performance of a multivariate analysis, whereby the data variables relating to the physical parameters are evaluated and model parameters are obtained. The model parameters describe the variation of the data variables. Thereby the samples are classified uniquely into classes. The identification of the classes is obtained when each sample is correlated to the characterisation information relating to said sample. Correlation in this respect is not necessarily a mathematical correlation. Correlation in this respect may also comprise the possibility of performing a comparison of data or fluorescence spectra.

15 Preferably, the model parameters are latent variables being weighted averages of the data variables.

The identification of the belonging to a class is obtained when the data variables of a sample are input to a trained classification system yielding either qualitative and/or quantitative information as to whether a sample belongs to a class.

20 In performing the data analysis it is often an advantage that the characterisation information is already available. Thereby it becomes possible to detect exactly those structures in the data that are relevant for detecting the difference between the classes and not just structures that may not be relevant for the classification.

25 The multivariate statistical methods suitable for the present invention are for example represented by chemometric methods like principal component analysis (PCA), partial least squares regression (PLS), soft independent modelling of class analogy (SIMCA) and principal variables (PV).

30 A non-exclusive list of other multivariate statistical methods include: Principal component analysis¹⁴, principal component regression¹⁴, factor analysis², partial least squares¹⁴, fuzzy clustering¹⁶, artificial neural networks⁶, parallel factor analysis⁴, Tucker models¹³, generalized rank annihilation method⁹, locally weighted regression¹⁵, ridge regression³, total least squares¹⁰, principal covariates regression⁷, Kohonen networks¹², linear or quadratic discriminant analysis¹¹, k-nearest neighbors

35

based on rank-reduced distances¹, multilinear regression methods⁵, soft independent modeling of class analogies⁸, robustified versions of the above and/or obvious non-linear versions such as one obtained by allowing for interactions or crossproducts of variables, exponential transformations etc.

5

The term "describing the variation of the data variables" means that the latent variables retain the relevant information regarding the variation, whereas "noise" is preferably not giving any significant part in the latent variables.

10 As an example of a multivariate data analysis technique, the use of principal component analysis – PCA – will be outlined [Jackson 1991] as this technique will be used in the following exemplary applications. An $I \times J$ data matrix, $X \in \mathbb{R}^{I \times J}$, is given where I is the number of rows (samples) and J is the number of columns. The number of variables will typically exceed the number of samples by far. This poses the
15 practical problem that the matrix is typically ill-conditioned. Thus, any traditional analysis using the whole set of raw data will lead to useless results due the numerical problems involved in handling the large amount of data

Using PCA, the original J variables are replaced by F ($\ll J$) *latent variables* which, in
20 this case, are also called principal components. These latent variables are found as weighted averages of the original variables in such a way that they provide the best possible description of the data in a least squares sense. Each latent variable consists of a score vector \mathbf{t} ($I \times 1$) and a loading vector \mathbf{p} ($J \times 1$). The loading vector is constrained to norm one and the score vector is found by regressing \mathbf{X} onto \mathbf{p}

25

$$\mathbf{t} = \mathbf{X}\mathbf{p}/\mathbf{p}^T\mathbf{p} = \mathbf{X}\mathbf{p}$$

For the first latent variable it holds that it minimizes

$$\|\mathbf{X} - \mathbf{t}\mathbf{p}^T\|_F^2$$

where $\|\bullet\|_F^2$ denotes the squared Frobenius norm. Thus, the first latent variable provides the least squares best-fitting rank-one model of \mathbf{X} . The second latent variable
30 is found under the constraint that the second score vector \mathbf{t}_2 is orthogonal to the first score \mathbf{t}_1 and that second loading vector \mathbf{p}_2 is orthogonal to the first loading \mathbf{p}_1 . Under this restriction, the second latent variable is found such that it provides the best possible fit to the data. Extracting F such components will yield a rank F model of the

data. Let the score matrix \mathbf{T} ($I \times F$) hold the score vectors \mathbf{t}_f , $f=1, \dots, F$ and the loading matrix \mathbf{P} ($J \times F$) holds the loading vectors \mathbf{p}_f , $f=1, \dots, F$ of this solution. It then holds that \mathbf{T} and \mathbf{P} provide the solution to

$$\arg \min_{\mathbf{G}, \mathbf{R}} \|\mathbf{X} - \mathbf{G}\mathbf{R}^T\|_F^2$$

5 and thus provide the best-fitting rank F solution. In practice, the solution to this problem can be found using a truncated singular value decomposition of \mathbf{X} . If \mathbf{U}_F holds the first F left singular vectors, \mathbf{V}_F holds the first F right singular vectors and \mathbf{S}_F is an $F \times F$ diagonal-matrix holding the first F singular values in its diagonal, then it holds that

10

$$\mathbf{T} = \mathbf{U}_F \mathbf{S}_F; \mathbf{P} = \mathbf{V}_F.$$

In order to choose the appropriate number of components, F , several strategies are possible. One approach is to use cross-validation [Wold 1978] in which elements are left out of the data in turn. For each set of elements left out, a model is fitted to the
15 remaining data and the model

$$\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}^T$$

is used to estimate the left out elements. After all elements have been left out once, the thus obtained residuals are used for calculating the predicted residual sum of squares (PRESS) and the number F for which PRESS is at its minimum is usually
20 taken to be the appropriate number of components. For exploratory purposes, it is usually sufficient to simply retain the first 2-5 components because these, per definition, retain most of the variation in \mathbf{X} . It is noted that if cross-validation is to be performed in this way, special algorithms have to be used because of the missing values in the data [Grung & Manne 1998].
25

The practical usefulness of PCA arises because of the information preserving compression of the data based on the empirical observations rather than on theoretical derivations. The scores \mathbf{T} can be seen as the coordinates of \mathbf{X} in the reduced space
30 defined by the truncated basis \mathbf{P} and the latent variables therefore provides a condensation of the original J variables into F new ones. This condensed representation is feasible because it allows a holistic visualization of the structure in the data and because it makes it possible to do quantitative analysis such as regression and classification in a straightforward way.

In some situations, the interest is to specifically make a quantitative model relating multivariate data to one or more responses by a regression model. This way, it is possible to measure future samples by the multivariate approach and then predict the response from the regression model. Such a regression problem suffers from the same problems as outlined above and for the same reason rank-reduced regression is often employed. As an example of such, the partial least squares regression – PLS – method will be described.

As before a multivariate set of data \mathbf{X} ($I \times J$) is available and further a response vector \mathbf{y} ($I \times 1$) is given. More responses can be handled as well, but this is not pursued here. The aim is to find a regression vector \mathbf{b} that provides a feasible solution to the regression problem

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

where \mathbf{e} ($I \times 1$) is a vector of unmodelled residuals. Using multiple linear regression ($J \leq I$) or similar approaches, it is possible to obtain a minimum variance unbiased estimate of \mathbf{b} but due to the constraint of being unbiased, the variance will, in practice, make the estimate useless for predictions in the situations considered here [de Jong 1995, Martens & Næs 1987, Wold et al. 1984]. Instead, a regression vector is sought that yields a low mean squared error, hence relaxing the restriction of being unbiased focusing on low total error. In PLS, this is achieved by extracting components sequentially such that each extracted score vector has maximal covariance with the yet unexplained variation in the response [Bro 1996, Martens & Næs 1989]. Usually \mathbf{X} and \mathbf{y} are centered by subtracting the column average from each column, thereby removing possible offsets. For centered \mathbf{X} and \mathbf{y} , the first component is determined by defining a weight vector as

$$\mathbf{w} = \frac{\mathbf{X}^T \mathbf{y}}{\|\mathbf{X}^T \mathbf{y}\|_F}$$

From this vector, the score vector \mathbf{t} is defined as

$$\mathbf{t} = \mathbf{X}\mathbf{w}$$

and finally a loading vector \mathbf{p} is defined as

$$\mathbf{p} = \mathbf{X}^T \mathbf{t} / \mathbf{t}^T \mathbf{t}$$

The rank-one model of \mathbf{X} is then given by $\mathbf{t}\mathbf{p}^T$ and the regression model relating the bilinear model of \mathbf{X} to \mathbf{y} is defined by the scalar

$$r = \mathbf{y}^T \mathbf{t} / \mathbf{t}^T \mathbf{t}.$$

5 giving the initial prediction of \mathbf{y} as $\mathbf{t}\mathbf{r}$. The model of \mathbf{X} ($\mathbf{t}\mathbf{p}^T$) and the prediction of \mathbf{y} ($\mathbf{t}\mathbf{r}$) are subtracted from \mathbf{X} and \mathbf{y} and the following component is determined similar to above but using the residuals of \mathbf{X} and \mathbf{y} as input. After calculating F components in this manner, the following matrices and vectors are given: \mathbf{T} ($I \times F$), \mathbf{P} ($J \times F$), \mathbf{W} ($J \times F$), \mathbf{r} ($F \times 1$). The regression vector can then be determined as

10

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^T \mathbf{W})^{-1} \mathbf{r}.$$

15

As for PCA, cross-validation is usually employed to determine the optimal rank of the model with the only difference being that whole samples are excluded in each cross-validation segment, and the residual error determined is the response error.

Other variables

20

The classification system may be obtained on the spectral information only. However, in some situations it may be appropriate to incorporate other variable(s) in the multivariate analysis.

25

These other variables may be variables relating to the sample supplying the spectral information or they may be variables that compensate for a specific condition of the sample.

30

Examples hereof may be the measurement of pH, electrolytes, temperature in the sample before subjecting it to spectroscopy, clinical parameters. Thereby variations in the other variables may be compensated for in the final classification.

35

In another embodiment other variables are variables relating to the animal, including a human being, to be characterised. Non-exclusive examples of these variables are hair colour, skin colour, age, sex, geographic origin, affiliation, prior diseases, hereditary background, medication intake, body conditions (such as e.g. surgery), stress level, medical diagnoses, subjective evaluations, and other diagnostic tests

(e.g. immunoassays, x-ray diagnosis, genomic information or an earlier chemometric test).

Pre-treatment

5

It is an advantage of the present invention that no pre-treatment of the sample is normally necessary.

10

However, for some of the samples or applications it may be necessary or convenient to perform an adjustment before subjecting the sample to spectroscopy.

15

Examples of pre-treatment may be adjustment of pH of the sample to a predetermined value, or heating or cooling the sample to a predetermined temperature. The sample may be treated with chemicals (complexing agents etc.) in order to develop e.g. fluorescent complexes involving inherent non-fluorescent molecules in the sample.

20

Other types of pre-treatment include addition of chemical substances, measurement under a gradient imposed by varying additions of chemical substances, simple chromatographic pre-treatments based on either chemical or physical separation principles.

Classification system

25

Another aspect of the present invention is the classification system for characterising a biological sample into at least one predetermined class.

30

When the classification system has been trained as discussed above, it is ready for classifying samples with unknown characteristics. The classification system preferably comprises the following components:

a) a sample domain for comprising a biological sample,

b) light means for exposing the sample to excitation light in the sample domain,

35

- c) a detecting means recording the physical parameter(s) of light emitted from the sample,
- d) optionally computing means for performing data handling of the physical parameters, obtaining data variables,
- e) optionally processing means for providing model parameters from data variables of the sample,
- f) at least one storage means for storing physical parameters and/or data variables and/or model parameters of the biological sample,
- g) at least one storage means for storing physical parameters and/or data variables and/or model parameters and characterisation information of a trained classification system,
- h) means for correlating physical parameters and/or data variables and/or model parameters from the sample with physical parameters and/or data variables and/or model parameters of the trained system, and
- i) means for displaying the characterisation class(es) of a sample.

The sample domain may be a sample chamber for accommodating a container with a liquid, a solid or a semi-solid sample. However, the sample domain may also be a domain in the individual to be classified in that the analysis can be performed on a sample in situ such as in the blood vessels or on the superficial body parts such as skin, nails, or hair.

The classification system may be provided as a whole unit, wherein the spectroscopy of the sample is conducted by the same unit from where the data relating to the characterisation classes of the sample is displayed.

It is however contemplated within the scope of the present invention, that the system is comprised of at least two units, wherein one unit is performing the steps a) to f), and another unit is performing the steps g) to i). Other units comprising other parts

of the system are also contemplated, such as one unit performing the steps a) to d) and storage means for storing physical parameters and/or data variables from f) or a) to g), and the other unit comprising the rest of the parts. Yet another unit comprises steps a) to c) in one unit and the remaining steps in the other unit.

5

By the system thus divided into at least two units, it is possible to obtain the spectroscopic information from a wide variety of decentral locations and perform the processing centrally. The data or the classification system may then be transmitted by any suitable means, such as conventional data transmission lines, for example the telephone lines, or via internet or intranet connections.

10

This facilitates the use of the classification system since any physician may provide the biological sample, have it subjected to spectroscopic analysis at his or her clinic and have the data correlated decentrally without the need of being capable of conducting this processing. The physician may then call the central unit to request the correlation and classification of the sample data. Depending on the transmission mode and equipment, the result may be displayed on a screen or printed on paper, or informed by telephone.

15

In addition to the result, other information may be provided, such as information regarding sample errors, for example the test requires a urine sample not a serum sample, information about the statistics, such as fuzziness, the degree of membership of a group, power and significance.

20

25 **Diagnosis**

In principle the classification system trained according to the invention may be used to characterise any biological sample with respect to any kind of information.

Interesting parts of the present invention relate to the possibilities of diagnosing a condition or the risk of acquiring a condition, such as a physical condition in an animal or a human being, from a spectrofluorimetric analysis of a biological sample from said animal or human being and relating the spectroscopic data with data in the classification system.

30
35

As for any other diagnostic tool, the present invention provides a diagnostic tool, that may give a strong indication of a disease or condition or a risk of such disease or condition, but for many of the diagnosis these may have to be confirmed by more specific diagnostic methods, more precisely directed to the specific diagnostic area.

5 However, due to the simplicity of the present invention, the precise diagnosis may be obtained faster and much more cost-effective than by hitherto known methods.

The disease detected may be any disease that provides combination of components in the biological sample that is detectable as a pattern by the fluorescence spectroscopy.

10

Thus the disease may be selected from any official disease classification system, such as ICD-9/10 (WHO's official international classification list), ICIDH-2 (International Classification of Functioning and Disability) but not limited to those two. Such classification system includes at least the following groups (the numbers in brackets refer to the ICD 9 list) of human diseases as well as similar diseases related to other animals:

15

Infectious and parasitic diseases (001-139)

20 Neoplasms (140-239)

Endocrine, nutritional and metabolic diseases, and immunity disorders (240-279)

Diseases of the blood and blood-forming organs (280-289)

Mental disorders (290-319)

Diseases of the nervous system and sense organs (320-389)

25 Diseases of the circulatory system (390-459)

Diseases of the respiratory system (460-519)

Diseases of the digestive system (520-579)

Diseases of the genitourinary system (580-629)

Complications of pregnancy, childbirth, and the puerperium (630-677)

30 Diseases of the skin and subcutaneous tissue (680-709)

Diseases of the musculoskeletal system and connective tissue (710-739)

Congenital anomalies (740-759)

Certain conditions originating in the perinatal period (760-779)

Injury and poisoning (800-999)

35 The sample may be classified to belong to a class for any of the diseases above,

quickly leading the examining physician to the most likely diagnosis. The sample classification may be confirmatory or it may have to be confirmed by more specific diagnostic tools.

- 5 Furthermore, the sample may be classified into more than one class, whereby a more refined diagnostic tool is provided.

10 For many of the diseases mentioned above, it is of utmost importance that an early diagnosis is obtained, but many of these diseases may be difficult to diagnose conventionally at the early stage due to very discrete symptoms. By the present method it is possible to get a clear indication of the disease at an early stage.

15 Furthermore, by the present invention it may furthermore be possible to reveal individuals susceptible to a specific disease, due to the classification of relevant biological samples from these individuals.

20 In particular in respect of cancer, the invention may be used to classify different forms of cancer, including cancer in various organs. Thus, the invention may be used to diagnose renal cancer from colon cancer for example. Furthermore, different stages of a cancer may be diagnosed, including precancerous stages. Furthermore different cancer aggressivity may be diagnosed for example the invention may identify high-risk cancer patients independently of whether the underlying cancer is anatomically localised in for example breast, lung or colon.

25 It is likewise conceivable that the present invention can be used for screening of individuals to identify those suffering from a particular disease or those being susceptible to a disease or those expected to suffer from the disease in the near future.

30 Also, the present invention may reveal individuals at risk due to environmental hazards, job environment or the like.

The present invention may also be used to diagnose a variety of abuse of medicine and/or narcotics, for example in relation to control, or in un-conscious or semi-conscious individuals that have to be treated for their abuse.

Another aspect of the invention may be to detect physical and/or psychological stress in an individual, and thereby detect persons at risk of acquiring stress related diseases.

- 5 Yet another aspect of the invention may be the detection of genetic modifications or inherited risk by examining a biological sample by fluorescence spectroscopy.

Yet another aspect of the invention may be to provide a quantitative answer to the degree of any of the above-mentioned situations (e.g. the amount of medicine, the
10 degree of risk etc.).

In a further embodiment the invention may be used as a tool for predicting the responsiveness to a specific treatment for an individual suffering from a particular disease. For example the invention may be used to classify individuals suffering from
15 cancer into classes of predicted responsiveness to chemotherapy, radiations and/or operation. Another example may be to predict the useful medication of depressive individuals.

Examples

20

Example 1

Example of the use of multidimensional sensorial fluorescence data analysis

25 Smokers/Non-smokers data

This example illustrates the usefulness of training a classification system on a small experimental data set. The data set comprises 18 samples taken from 14 individuals.

Sampling and measurements

30 Urine was collected from several male persons and measured spectrofluorimetrically. Approximately half of the testees were smokers, the other half non-smokers. Some samples were measured the same day, others up to 4 days after sampling. The samples were kept at -4°C until measurement was performed. The urine was

not diluted before measurement. Fluorescence spectra were recorded on a Perkin-Elmer LS-50B spectrofluorimeter using front face illumination.

5 Scanning was performed from excitation wavelengths 230 nm to 500 nm and from emission wavelengths 268 nm to 900 nm.

Data handling

For each thus obtained fluorescence landscape, the obvious non-bilinear parts (emission below excitation and zero- and first-order Rayleigh scatter) were removed and the corresponding elements denoted 'missing'. This lead to data of the type shown in Figure 1.

For each sample, i , a matrix X_i is thus obtained of size $J \times K$ with J being the number of emission wavelengths and K being the number of excitation wavelengths. The whole data set is arranged in a three-way tensorial structure with typical elements x_{ijk} , $i = 1, \dots, I$, $j = 1, \dots, J$, $k = 1, \dots, K$. This three-way structure may, geometrically, be interpreted as a box of data, where each horizontal slice corresponds to a specific sample, each vertical slice corresponds to a specific emission wavelength, and each frontal slice corresponds to a specific excitation wavelength.

20 In the following, this three-way tensorial array is matricized, i.e. rearranged into a two-way matrix called Z where each row corresponds to a sample, and holds all combinations of excitation and emission. In this setup - interpreted as in ordinary multivariate data analysis - there are 4003 variables (plus a fraction that is removed because it contains either variables which are set to missing or variables with extremely small variance).

25 Data modelling

This data matrix is subjected (Matlab, version 5.2) to principal component analysis, in which the 4003 data variables are replaced with three *latent* variables. These latent variables are weighted averages of the original data variables, defined so that the projection of Z onto the space spanned by these, retain as much variation as possible. That is, the latent variables T of size 18×3 are defined through a set of weights, P (4003×3) as the solution to

$$\max_P \|ZP(P^T P)^{-1} P^T\|_F^2$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Because the weight matrix is chosen to be orthogonal (without loss of generality), this expression can be reduced to

$$\max_P \|ZPP^T\|_F^2$$

From this expression, the latent variables **T** are found as the coordinates of **Z** in the reduced space defined by the truncated basis **P**

$$\mathbf{T} = \mathbf{ZP}.$$

Note, that the latent variables are found using *no* information about the status of the persons (smoker or non-smoker). The latent variables provides a condensed picture of the original 4003 variables, and a simple graphical representation of these is sufficient to illustrate the power of this compression (Figure 2).

As can be seen in the plot, all non-smokers fall above the dashed line and all but one smoker falls below the dashed line. Disregarding for now the one outstanding sample (S3), it is easily seen that this simple plot provides a powerful tool for assessing whether a person is a smoker or not. Simply by measuring the fluorescence excitation-emission data from a urine sample, and projecting the obtained data onto the factor weights found above, a set of scorings on the three latent variables is obtained for a new person. When these are plotted in the above plot, the position of the sample above or below the dashed line enables an assessment of whether the person is a smoker or not. In fact, the sample of the person identified as S2 in the lower right part, was left of the initial analysis and only positioned after the plot was generated. As can be seen the position is correctly below the dashed line, indicating that the person is a smoker. This graphical assessment can be automated in a number of ways using appropriate pattern recognition techniques.

The one smoker-sample located above the dashed line indicates an erroneous condition. However, in this initial feasibility study, no detailed information on the individual persons were available nor of their smoking habits. Hence, there can be numerous reasons for this particular position, such as the person had not been smoking that particular day and the day before etc.

Example 2

Detecting fasting condition

5

A patient undergoing a surgical procedure in general anesthesia is exposed to a relatively high risk if he or she has been eating or drinking before becoming intubated. Under a planned procedure the patients have been instructed not to ingest before the surgical procedure. However, the patients (especially children or elderly persons) will not always refrain from ingesting. In both the planned and acute procedure it is of help for the physician to know whether a patient has been eating or not. Thus, such a test will be a feasible 'add-on' to other tests with low marginal cost.

Samples and solutions

15

This study includes 9 normal persons fasting and 14 normal persons not fasting. The conditions for all 23 persons were identical except for the question of whether the persons had been fasting or not.

20

For each person a blood sample was taken and blood plasma therefrom frozen. The blood plasma samples were defrosted and measured at room temperature. The samples were measured undiluted front face in a 1 mm cuvette on a Perkin Elmer LS50B (Copenhagen University). The excitation wavelength interval range was 230-400 nm (10 nm steps) and the excitation and emission slits were 4 nm and 3 nm, respectively. The scan rate was 1000 nm/min. In all, a total of 3834 variables were measured for each sample. At each excitation wavelength, the emission was removed in the range from 250 to 22 nm above the excitation wavelength in order to remove Rayleigh scatter and other irrelevant phenomena [Bro 1998, Bro 1999]. Upon removal of these variables, a total of 2020 variables were retained. A typical landscape is shown in Figure 3.

25

30

Results

The data were fitted by a PCA model. The model indicated that at least up to six components contained valid information. For the present purpose, it means that the main systematic part of the fluorescence variation is retained in these six new variables. A resulting score plot is shown below (Figure 4). It is a three-dimensional

scatter plot of score one, two and five. Each plot represents the relative position of one person with respect to that persons fluorescence fingerprint. It is immediately seen in the plot that all fasting persons appear to the right in the plot and all non-fasting persons appear to the left.

5

The significance of this particular plot can be described as follows. In estimating the six components in the PCA model, no use whatsoever has been made of the fasting-information. The PCA model is only based on the fluorescence data. The empirical observation that it is possible to assign areas of the plot to only fasting persons and areas to only non-fasting persons means that a discrimination between the two groups has been achieved. Thus, for a person where it is unknown whether the person is fasting, it is possible to measure a corresponding fluorescence landscape under similar conditions and thereby obtain the scores for that particular person. Inserting these scores in the plot above, it is then possible to evaluate or verify whether the person is fasting or not by simply monitoring to which side of the indicated line the point is positioned.

10

15

20

More elaborate decision rules can easily be envisioned using e.g. linear discriminant analysis [Indahl et al. 1999], SIMCA [Wold & Dunn, III 1983] or some similar classification approach. However, for this feasibility study it suffices to show that discrimination is possible to achieve with multivariate analysis of fluorescence landscapes.

Example 3

25

Analysis of colon cancer data

Having a simple tool for detecting colon cancer is a very interesting application of the current invention.

Materials and methods

30

The data gathered here, include 77 samples (9 normal persons; 13 with benign tumor; 11 Dukes A; 14 Dukes B; 15 Dukes C; 15 Dukes D). For each person a blood sample was taken and blood plasma therefrom frozen. The blood plasma samples were defrosted and measured at room temperature. The samples were measured undiluted front face in a 1 mm cuvette on a Perkin Elmer LS50B (Copenhagen Uni-

versity). The excitation wavelength interval range was 230-400 nm (10 nm steps) and the excitation and emission slits were 4 nm and 3 nm, respectively. The scan rate was 1000 nm/min. In all, a total of 3834 variables were measured for each sample. At each excitation wavelength, the emission was removed in the range from
5 250 to 22 nm above the excitation wavelength in order to remove Rayleigh scatter and other irrelevant phenomena [Bro 1998, Bro 1999]. Upon removal of these variables, a total of 2020 variables were retained.

Results

For each class of samples, a principal component analysis (PCA) model is fitted to
10 the fluorescence data. This is important for exploring the homogeneity of the group and for eliminating obvious erroneous samples.

As an example, a PCA model of the data of persons with benign tumors is discussed. In this group, there seems to be several individuals located distinctly isolated (4490, 4499, 8319, 4506) in the score plots (Fig 5). The remaining persons are
15 situated in the same group in the score plot. The reasons for the behavior of the outlying samples can be related to the patients (extreme patients in some sense), to the sampling of the blood (extreme sampling in some sense) or to the actual measurements. For e.g. 4490 the person was later found to be incorrectly classified as benign. For 4506 the technician noted that the suspension was cloudy indicating
20 incorrect treatment of the sample. For 8319 the sample was noted to have precipitated matter, whereas for 4499 no reason was found for its behavior, besides the outlying behavior of the measured fluorescence data. In order to assure that the subsequent results are as robust as possible, the four samples were excluded; the
25 three because of erroneous sample treatment or measurement and the fourth because of assumed but unknown erroneous sample treatment or measurement.

Similar outlying samples were found in other groups as well. E.g. for the group of Dukes A one sample had a very different fluorescence pattern and was excluded.
30 For the Dukes B samples, four such samples were observed. For Dukes C only sample VB106 is moderately outlying. For Dukes D, sample VB76 is moderately outlying. All in all, 11 samples out of the 77 were excluded. An explanation of the erroneous sample treatment or measurement was found for most of these samples, which makes the decision to exclude these valid and reasonable. It must however,

be borne in mind, that for five of the samples, there was no explanation found for the strange behavior. Hence, excluding these from the subsequent classification model is somewhat hazardous, because similar correct samples might be anticipated in real applications. Nevertheless, for the present feasibility study, the samples are
 5 considered as outliers of which the cause is presently unknown.

In order to quantify how well these data can be used for screening for cancer, the data were split up into two groups: Persons without cancer (18) and persons with cancer (48). A cross-validation was performed in the following way. One person was
 10 left out in turn and subsequently a PCA model was fitted to each group of data. Thus two PCA models were built. For the non-cancer group, five components were used and for the cancer group seven components were used. The data from the left-out person was subsequently fitted to the two independently obtained models yielding 1) a set of score values for the sample and 2) a set of residuals of fluorescence varia-
 15 tion of that sample that the model could not explain. For one model, the score values of the new sample, \mathbf{t} ($1 \times F$) and the scores from the calibration data \mathbf{T} ($I \times F$) are used for calculating the T^2 statistic as $T^2 = \mathbf{t}(\mathbf{T}^T \mathbf{T})^{-1} \mathbf{t}$ and the Q statistic as $\mathbf{e}^T \mathbf{e}$ where \mathbf{e} ($J \times 1$) is the vector of residual variation in the fluorescence data not explained by the model. The ratio of these values and the corresponding confidence limits ob-
 20 tained from the model are calculated (hence a value above one indicates that the sample is different). These two ratios are squared and summed and the square-root is used to test for class belongingness. If this number is less than the square-root of two, the sample is assigned to the class. If the sample is assigned to both classes, the one with the smallest number is the one chosen.

25

Using this approach the following classification result is obtained.

Table 1. Classification results

	Normal	Cancer
Total	18	48
Correctly classified	6	48
Incorrectly classified	12	0

30 It is observed that 82% are correctly classified and no false negatives are obtained.

Example 4

Fluorescence measurements of urine from cardiac patients

5 This example illustrates the treatment of outlying data and the ability of the invention to classify patients according to cardiac problems.

Samples

Eight urine samples were collected from seven men and one woman (post-menopause) who all were diagnosed with angina pectoris (samples #1-8). No other information was available from these patients. For comparison, urine
10 samples were collected from five, arbitrarily chosen men (samples #9-13).

Measurements

Excitation-emission matrices were measured on the undiluted samples in a cuvette with 2 mm light path using front-face geometry on a Perkin-Elmer LS50B spectrofluorometer. In 28 consecutive scans, the excitation wave-
15 length was shifted in 10 nm steps from 230 to 500 nm. Emission intensity was recorded starting 20 nm after the excitation wavelength until two times the excitation wavelength minus offset (or 900 nm). Thus, neither first nor second order Rayleigh scatter were recorded. Emission intensity was measured in intervals of 0.5 nm. Spectral bandwidth on both monochromators
20 was 5 nm. Scan-rate was 500 nm/min.

In total, fluorescence intensity was measured at 17828 different combinations of excitation and emission wavelengths and the values exported to Matlab, version 5.2.1.

Results

25 Although the data have a three-way structure (samples \times excitation wavelength \times emission wavelength) this is disregarded and the data are rearranged to a two-way structure (samples \times combination of excitation and emission wavelength) as illustrated in Figure 6. In this the 28 spectra, being
30 arranged successively on an increasing wavelength scale, are displayed in

an overlay fashion. The thus obtained two-matrix is centered by subtracting from each column its average value. By means of Principal Component Analysis this centered matrix is modelled by three principal components obtained from a singular value decomposition of the centered matrix.

5

An initial analysis reveals that patient #5 is quite extreme as compared to the remaining patients. This is illustrated in Figure 7 in a so-called influence plot. In a two-component model patient five has a very large residual variation (upper left corner) whereas in a three-component model patient five has an extremely high leverage (lower right corner). This result shows that after two components most data are well described except for patient five. Consequently, in the third component the fifth patient gets a high leverage, which means that this patient is determining this component. This is a typical example of an extreme outlier. If the cause for the extreme behaviour is instrumental, the patient's data must be excluded as an incorrect measurement. If the cause is biological diversity, this diversity must be better represented in the data by incorporating more similar samples. As there are no further data in this specific investigation, the only suitable procedure is to exclude this sample.

Indirectly, the appearance of the outlying sample is an important illustration of one of the very important benefits of using exploratory data-analysis and having many physical parameters at disposal. Had it not been possible to detect the outlying sample, conclusions from the analysis could have been misleading. The model would be reflecting the difference between the samples as such and the extreme sample five, rather than explaining the inter-differences and patterns between *all* samples. The availability and use of this evaluating tool during the model-building step shows that quality-deteriorating samples can be excluded, thus leading to improved models with improved validity.

30

Refitting the principal component model without sample five, a score-plot is obtained as shown in Figure 8. Only the first two score vectors, PC1 and PC2, are displayed. The most important *latent* variables, PC1 and PC2 rep-

resent 97% of the original variation in the fluorescence data obtained in this investigation. The samples separate into two distinct clusters: Those below the dashed line in the lower left corner all represent samples from persons diagnosed with angina pectoris while the samples in the upper right corner all
5 represent persons that are not diagnosed angina pectoris. Thus, it is clearly possible to separate diseased from healthy persons based on the fluorescence data alone.

It is indeed a significant finding, that the fluorescence data so clearly separate the two groups. Importantly, it is not merely the intensity of fluorescence
10 that separates the patients. Differences in intensities are normally reflected in the first principal component of spectral data. In this case, however, the second component is also important for obtaining separation. In fact, the third component - not shown - also helps in obtaining further separation. This result indicates, that more subtle spectral components can be increasingly
15 helpful in the discrimination between the groups.

As a larger data set becomes available the procedure outlined here will easily be formalised into a classification model that can identify persons with cardiac problems within a population of otherwise healthy individuals.
20

Example 5

Fluorescence measurements of urine samples with added bacteria

25

The purpose of the example was to investigate if fluorescence spectra measured directly on urine samples correlate with different added levels of bacteria in the urine.

Samples

30 Seven urine samples spiked with 10^2 to 10^8 E. coli bacteria pr. ml and a control sample with no bacteria added. The eight samples were delivered by Alice Friis-Møller, Hvidovre Hospital and kept in a freezer until measurement.

Measurements

The samples were measured at front face at room temperature in a 2 mm cuvette on a Perkin Elmer LS50B (Copenhagen University). The excitation wavelength interval range was 230–400 nm (10 nm steps) and the excitation and emission slits were 4 nm and 3 nm, respectively. The scan rate was 1000 nm/min. The data were imported to Matlab using every 5th emission wavelength giving a step of 2.5 nm in the emission scans. An important note is that the samples were measured in a sequence corresponding to the increase in bacteria content.

Results

10 Raw data

The raw unfolded data are averaged with a factor 10 over the variables so the matrix dimensions become 8 samples x 384 variables. This corresponds to circa 25 nm steps in the emission scans. Fig. 9.

PCA

15 A PCA is performed on both the mean centered and the auto scaled unfolded spectral data. Variables with standard deviation of 0 and variables including missing values (NaNs) are excluded.

In Figure 10 and 11 the score plots from these models are shown. No large differences are seen between the auto scaled and the mean centered models.

20

PC1 clearly reflects the increase in bacteria content.

PLS models

Mean centered PLS models with fluorescence spectra as the independent variables and the logarithm of the bacteria content (10^2 to 10^8) as the dependent variable are developed.

25

It is observed that there is a strongly non-linear relationship between the spectra and the number of bacteria cells, but the function is monotone. (Fig. 12)

30 Conclusions

There seems to be a non-linear relationship between the spectra and the number of bacteria cells. This can be corrected for by some non-linear transformation of e.g.

the dependent variable. Furthermore, it is important to note that the measurement order is crucial and should be randomised in a follow up study.

Example 6

5

Basic fluorescence measurements of blood plasma

The purpose of the example was to investigate the effect of dilution and pH on the fluorescence spectra measured on blood plasma. Both transmission and front face sample presentations were investigated.

10

Samples

A 10 ml pool of blood plasma samples from Hvidovre Hospital was produced. Samples were taken from this pool and diluted with 0.9% sterile NaCl, and the following dilutions were performed for front face:

15	1:2	3.5 ml pool	+	3.5 ml NaCl (A)
	1:5	2.0 ml A	+	3.0 ml NaCl (B)
	1:10	1.0 ml A	+	4.0 ml NaCl (C)
	1:25	2.0 ml C	+	3.0 ml NaCl
	1:50	0.5 ml B	+	4.5 ml NaCl
20	1:100	1.0 ml pool	+	99.0ml NaCl (F)
	1:200	2.5 ml F	+	2.5 ml NaCl
	1:500	1.0 ml F	+	4.0 ml NaCl
	1:700	1.0 ml F	+	6.0 ml NaCl
	1:2000	0.25ml F	+	4.75ml NaCl
25	1:3000	0.20ml F	+	5.8 ml NaCl
	1:5000	0.10ml F	+	4.9 ml NaCl

and for transmission:

	1:2	2.8 ml pool	+	2.8 ml NaCl (A)
	1:5	2.0 ml A	+	3.0 ml NaCl (B)
30	1:10	1.0 ml A	+	4.0 ml NaCl (C)
	1:25	2.0 ml C	+	3.0 ml NaCl
	1:50	0.5 ml B	+	4.5 ml NaCl
	1:100	0.25 ml pool	+	24.75ml NaCl (F)
	1:200	2.5 ml F	+	2.5 ml NaCl

	1:500	1.0 ml F +	4.0 ml NaCl
	1:700	1.0 ml F +	6.0 ml NaCl
	1:1000	0.3 ml F +	2.7 ml NaCl
	1:2000	0.25ml F +	4.75ml NaCl
5	1:3000	0.20ml F +	5.8 ml NaCl
	1:5000	0.10ml F +	4.9 ml NaCl

Buffers were produced as follows:

pH circa 9: 0.1 M NaH_2PO_4 (1.785g NaH_2PO_4 to 100ml H_2O).

pH circa 4: 0.1 M Na_2HPO_4 (1.382g Na_2HPO_4 to 100ml H_2O).

10 pH circa 7: 1.379g NaH_2PO_4 + 1.787g Na_2HPO_4 to 100ml H_2O .

pH circa 1: 0.1 M HCl (0.81 ml concentrated HCl to 100ml H_2O).

and dilutions for the buffer experiment were:

	1:2	2.0 ml pool +	2.0 ml buffer
	1:10	0.4 ml pool +	3.6 ml buffer (B)
15	1:200	0.2 ml B +	3.8 ml buffer
	1:1000	0.1 ml B +	9.9 ml buffer

All possible combinations of pH levels and dilutions were measured resulting in 16 (4x4) spectral landscapes. Both front face and transmission were tested.

Measurements

- 20 The samples were defrosted and measured at room temperature. The samples were measured in a standard 10x10 mm cuvette on a Perkin Elmer LS50B. The excitation wavelength interval range was 230-400 nm (10 nm steps) and the excitation and emission slits were 4 nm and 3 nm, respectively. The scan rate was 1000 nm/min. The data were imported to Matlab using every 5th emission wavelength giving a step of 2.5 nm in the emission scans.
- 25

Results

Experiment 1: Dilution of blood samples measured in front face mode

Raw data

- 30 In Figure 13 examples of front face fluorescence spectra of an undiluted sample and the same sample diluted 1:5000 are shown. Very different spectral signals both with respect to intensity and shape are obtained for the two samples.

At low excitation wavelengths the signal intensities at first increase with dilution (probably due to quenching) followed by a decrease in signal intensity by further dilution. At high excitation wavelengths a (almost linear) decrease in signal intensity is seen with dilution. This is illustrated in Figures 14 and 15.

5

PCA

A PCA is performed on the mean centered unfolded spectral data. Variables with standard deviation of 0 and variables including missing values (NaNs) are excluded. Systematic score patterns are seen for up to 5 to 6 PCs and Figure 16 shows the scores for 1 to 4 PCs. Variance explained is 79.38%, 16.32%, 2.60%, 1.45% and 0.24% for the first 5 PCs, respectively.

10

Experiment 2: Dilution of blood samples measured in transmission mode

Raw data

In Figure 17 examples of transmission fluorescence spectra of an undiluted sample and the same sample diluted 1:5000 are shown. Again, very different spectral signals both with respect to intensity and shape are obtained for the two samples.

15

At low excitation wavelengths the signal intensities at first increase with dilution (probably due to quenching) followed by a decrease in signal intensity by further dilution. At high excitation wavelengths a (almost linear) decrease in signal intensity is seen with dilution. This is illustrated in Figures 18 and 19.

20

PCA

A PCA is performed on the mean centered unfolded spectral data. Variables with standard deviation of 0 and variables including missing values (NaNs) are excluded. Systematic score patterns are seen for up to 5 to 6 PCs as for front face mode and Figure 20 shows the scores for 1 to 4 PCs. Variance explained is 85.43%, 11.33%, 1.91%, 0.99% and 0.30% for the first 5 PCs, respectively.

25

30

PLS models

Mean centered PLS models with fluorescence spectra as the independent variables and the dilution factor as the dependent variable are developed for both front face and transmission mode measurements.

It is observed that there is a linear relationship from dilution factor 1:25 to 1:5000 for front face and for 1:50 to 1:5000. It seems that the relationship is non-linear below these dilutions factors. PLS modelling was also tested with auto scaled data to give the high excitation wavelengths more influence. Equal results were obtained although the linear range of the models could be expanded to approx. 1:10. PLS modelling was also tested with log(dilution factor) and nice linear models over the whole range were obtained. Only the undiluted sample seemed to deviate a little.

Experiment 3: Effect of pH & dilution on the measured spectra

PCA

PCA is performed on each of the data sets recorded in transmission and front face mode. A score plot from a PCA on all samples is shown in Figure 23.

No huge differences are seen with respect to pH levels, see also Figure 24.

Conclusions

It is important to measure at least two (or even better three or four) different dilutions of the blood plasma samples: the undiluted sample and the same sample diluted 1:2 in front face and diluted 1:200/1:100 in transmission. Not surprisingly, front face intensities are higher for samples measured at no or low dilution factors, while the opposite holds for samples with high dilution factors. Note, that it is possible to measure in transmission mode on the undiluted sample. The tested pH levels do not seem have large effects on the measured spectral shapes or intensities.

Reference List

1. Alsberg BK, Goodacre R, Rowland JJ, Kell DB, Classification of pyrolysis mass spectra by fuzzy multivariate rule induction-comparison with regression, k-nearest neighbour, neural and decision-tree methods, *Analytica Chimica Acta*, 1997, **348**, 389-407.

2. Bartholomew DJ, The foundation of factor analysis, *Biometrika*, 1984, **71**, 221-232.
3. Björkström A, Sundberg R, A generalized view on continuum regression, *Scandinavian Journal of Statistics*, 1999, **26**, 17-30.
- 5 4. Bro R, PARAFAC. Tutorial and applications, *Chemom Intell Lab Syst*, 1997, **38**, 149-171.
5. Bro R, Multiway calibration. Multi-linear PLS, *Journal of Chemometrics*, 1996, **10**, 47-61.
- 10 6. Bro R, Multi-way Analysis in the Food Industry. Models, Algorithms, and Applications. Ph.D. thesis, University of Amsterdam (NL), 1998,
7. Bro R, Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis, *Chemom Intell Lab Syst*, 1999, **46**, 133-147.
- 15 8. Cheng B, Titterton DM, Neural Networks: A Review from a Statistical Perspective, *Statistical Science*, 1994, **9**, 2-54.
9. de Jong S, Kiers HAL, Principal covariates regression. Part 1. Theory, *Chemom Intell Lab Syst*, 1992, **14**, 155-164.
- 20 10. Esbensen K, Wold S, SIMCA, MACUP, SELPLS, GDAM, SPACE & UNFOLD: The way towards regionalized principal components analysis and subconstrained N-way decomposition - with geological illustrations, *Proc Nord Symp Appl Statist*, Stavanger, 1983,
- 25 11. Faber NM, Buydens LMC, Kateman G, Generalized rank annihilation method. I: Derivation of eigenvalue problems, *Journal of Chemometrics*, 1994, **8**, 147-154.
12. Golub GH, Hansen PC, O'leary D, Tikhonov Regularization and Total Least Squares, *SIAM Journal of Numerical Analysis*, 1999, **21**, 185-194.
13. Grung B, Manne R, Missing values in principal component analysis, *Chemom Intell Lab Syst*, 1998, **42**, 125-139

14. Indahl UG, Sahni NS, Kirkhus B, Næs T, Multivariate strategies for classification based on NIR-spectra - with application to mayonnaise, *Chemom Intell Lab Syst*, 1999, **49**, 19-31.
15. Jackson J, *A User's Guide to Principal Components*. Wiley & Sons, New York, 1991
16. Kohonen T, Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, 1982, **43**, 59-69.
17. Kruskal JB, Harshman RA, Lundy ME, Some relationships between Tucker's three-mode factor analysis and PARAFAC/CANDECOMP. 1983,
18. Martens H, Næs T, *Multivariate calibration*. John Wiley & Sons, Chichester, 1989, Martens H, Næs T, *Multivariate calibration by data compression, Near Infrared Technology in the Agricultural and Food Industries*, (Eds. Williams, P and Norris, K), The american association of cereal chemists, Inc., St. Paul, 1987, 57-87.
19. Næs T, Isaksson T, Some modifications of locally weighted regression (LWR), *NIR news*, 1994, **5**, 8-9.
20. 16. Rajko R, Treatment of model error in calibration by robust and fuzzy procedures, *Analytical Letters*, 1994, **27**, 215-228.
21. 17. Wold S, Dunn WJ, III, Multivariate quantitative structure-activity relationships (QSAR): conditions for their applicability, *J Chem Inf Comput Sci*, 1983, **23**, 6-13.
22. Wold S, Cross-validatory estimation of the number of components in factor and principal components models, *Technometrics*, 1978, **20**, 397-405.
23. Wold S, Albano C, Dunn WJ, III, Edlund U, Esbensen KH, Geladi P, Hellberg S, Johansson E, Lindberg W, Sjöström M, *Multivariate data analysis in chemistry, Chemometrics. Mathematics and Statistics in Chemistry*, (Ed. Kowalski, BR), D. Reidel Publishing Company, Dordrecht, 1984, 17-95.

Claims:

- 5 1. A method of training a classification system for characterising a biological sample with respect to at least one condition, comprising
- 10 a) obtaining a biological sample from an animal, including a human, wherein said biological sample is selected from body fluids and/or tissue, wherein the tissue sample is not associated with said condition(s),
- 15 b) obtaining characterisation information related to each biological sample,
- c) exposing the sample to excitation light within a predetermined range of wavelength,
- 20 d) determining physical parameter(s) of light emitted from the sample,
- e) repeating step a) to d) until the physical parameters of all training samples have been determined,
- 25 f) optionally performing a data handling of the obtained physical parameters obtaining data variables,
- g) optionally performing a multivariate data analysis of the data variables and optionally of characterisation information obtaining model parameters describing the variation of the data variables,
- 30 h) classifying the biological samples into at least two different classes correlated to the characterisation information, obtaining a trained classification system.
2. The method according to claim 1, whereby step g) further comprises selection of latent variables being weighted averages of data variables.

3. The method according to claim 1, wherein the biological sample is selected from blood, serum, plasma, saliva, urine, milk, cerebrospinal fluid, tears, nasal secrete, semen, bile, lymph, sweat and/or faeces.
- 5 4. The method according to claim 1, wherein the biological sample is a tissue sample.
5. The method according to claim 4, wherein the tissue sample is a biopsy of tissue selected from muscle, cutis, subcutis, kidney, brain, and liver or a sample of hair
10 or nails.
6. The method according to claim 3, wherein the biological sample is urine, milk, blood, plasma or serum.
- 15 7. The method according to claim 1, wherein the wavelength of the excitation light is in the range of from 100 nm to 1000 nm, such as from 100 to 800 nm.
8. The method according to claim 7, wherein the wavelength of the excitation light is in the range of from 200 nm to 800 nm, such as from 200 nm to 600 nm.
20
9. The method according to claim 1, wherein the physical parameter determined is selected from fluorescence intensity, fluorescence lifetime, phosphorescence intensity, phosphorescence lifetime, polarisation, polarisation lifetime, anisotropy, anisotropy lifetime, phase-resolved emission, circularly polarised fluorescence, fluorescence-detected circular dichroism, and any time dependence of
25 the two last mentioned parameters.
10. The method according to claim 1, wherein the spectral distribution of light emitted ranging from 200 nm to 800 nm is generated.
30
11. The method according to claim 2, wherein the ratio of number of training samples to the expected number of latent variables is at least 5:1, preferably at least 10:1.

12. The method according to claim 1, wherein the multivariate data analysis is selected from: Principal component analysis, principal component regression, factor analysis, partial least squares, fuzzy clustering, artificial neural networks, parallel factor analysis, Tucker models, generalised rank annihilation method, locally weighted regression, ridge regression, total least squares, principal co-
5 variates regression, Kohonen networks, linear or quadratic discriminant analysis, k-nearest neighbours based on rank-reduced distances, multilinear regression methods, soft independent modelling of class analogies, robustified versions of the above and/or obvious non-linear versions such as one obtained by allowing
10 for interactions or crossproducts of variables, exponential transformations etc.
13. The method according to claim 1, wherein the data handling of step f) is selected from a one-way matrix of spectral information, a two-way matrix of spectral information, a three-way matrix of spectral information, a four-way matrix of spectral
15 information and, a five-way or higher order matrix of spectral information.
14. The method according to claim 1, wherein other variable(s) is included in the multivariate analysis of step g).
- 20 15. The method according to claim 14, wherein the other variable(s) is selected from a pH value of the sample, concentration of various electrolytes in the sample, concentration of any other relevant compound in the sample, temperature, chemical parameters or any other physical property of the sample.
- 25 16. The method according to claim 1, wherein other variable(s) related to the animal, including a human being, is included in the multivariate analysis of step g)
- 30 17. The method according to claim 16, wherein the other variable(s) is selected from any parameter relating to the bodily or mental condition, hair colour, skin colour, age, sex, geographic origin, affiliation, hereditary background, stress level, medical diagnosis, subjective evaluations or clinical parameters.
18. The method according to claim 1, wherein the sample is pre-treated before subjecting the sample to step c).

19. The method according to claim 18, wherein the pre-treatment comprises adjustment of pH of the sample to a predetermined value.
20. The method according to claim 1, wherein a classification system for diagnostic purposes with relation to heart diseases is obtained.
21. The method according to claim 1, wherein a classification system for diagnostic purposes with relation to abuse of medicine or narcotics is obtained.
22. A diagnostic classification system comprising
- a) a sample domain for comprising a biological sample,
 - b) light means for exposing the sample to excitation light in the sample domain,
 - c) a detecting means recording the physical parameter(s) of light emitted from the sample,
 - d) optionally computing means for performing data handling of the physical parameters, obtaining data variables,
 - e) optionally processing means for providing model parameters from data variables of the sample,
 - f) at least one storage means for storing physical parameters and/or data variables and/or model parameters of the biological sample,
 - g) at least one storage means for storing physical parameters and/or data variables and/or model parameters and characterisation information of a trained classification system,
 - h) means for correlating physical parameters and/or data variables and/or model parameters from the sample with physical parameters and/or data variables and/or model parameters of the trained system, and

i) means for displaying the characterisation class(es) of a sample.

23. The system according to claim 22, wherein the model parameters are latent variables being weighted averages of the data variables.

5

24. The system according to claim 22, wherein the biological sample is a liquid sample, such as a sample selected from blood, serum, saliva, milk, urine, cerebrospinal fluid, tears, nasal secrete, semen, bile, lymph, sweat and/or faeces.

10

25. The system according to claim 22, wherein the biological sample is a tissue sample.

26. The system according to claim 25, wherein the tissue sample is a biopsy of tissue selected from muscle, cutis, subcutis, kidney, brain, and liver or a sample of hair or nails.

15

27. The system according to claim 24, wherein the biological sample is urine, milk, blood, plasma or serum.

20

28. The system according to claim 22, wherein the light means is arranged to emit light having a wavelength in the range of from 100 nm to 1000 nm, such as from 100 to 800 nm.

25

29. The system according to claim 28, wherein the light means is arranged to emit light having a wavelength in the range of from 200 nm to 800 nm, such as from 200 nm to 600 nm.

30

30. The system according to claim 22, wherein the physical parameter determined is selected from fluorescence intensity, fluorescence lifetime, phosphorescence intensity, phosphorescence lifetime, polarisation, polarisation lifetime, anisotropy, anisotropy lifetime, phase-resolved emission, circularly polarised fluorescence, fluorescence-detected circular dichroism, and any time dependence of the two last mentioned parameters.

31. The system according to claim 22, wherein the detecting means is selected from a photomultiplier, a scanning camera, for example a vidicon, a CCD camera, a CMOS, or a diode array.
- 5 32. The system according to claim 22, being divided into at least a first unit and a second unit, wherein said first unit comprises the parts a) to at least c) of the system, and the second unit comprises the other parts.
- 10 33. The system according to claim 22, further including means for measuring other variable(s) of the sample.
34. The system according to claim 33, wherein the other variable(s) is selected from a pH value of the sample, concentration of various electrolytes in the sample, concentration of any other relevant compound in the sample, temperature, chemical parameters or any other physical property of the sample.
- 15 35. The system according to claim 22, further including means for entering other variables.
- 20 36. The system according to claim 35, wherein the other variable(s) is selected from any parameter relating to the bodily or mental condition, hair colour, skin colour, age, sex, geographic origin, affiliation, hereditary background, stress level, medical diagnosis, subjective evaluations or clinical parameters.
- 25 37. The system according to claim 22, wherein the sample is pre-treated before subjecting the sample to step b).
38. The system according to claim 37, wherein the pre-treatment comprises adjustment of pH of the sample to a predetermined value.
- 30 39. The system according to claim 22, being a classification system for diagnostic purposes with relation to heart diseases.
40. The system according to claim 22, being a classification system for diagnostic purposes with relation to abuse of medicine or narcotics.
- 35

41. A method for characterising a biological sample of an animal, including a human, comprising
- 5 a) obtaining a biological sample from the animal or human,
 - b) exposing the sample to excitation light,
 - c) determining the physical parameter(s) of light emitted from the sample,
 - 10 d) optionally performing a data handling of the obtained physical parameters obtaining data variables,
 - e) storing the physical parameters and/or data variables and/or model pa-
15 rameters,
 - f) optionally providing model parameters from data variables of the sample,
 - g) obtaining physical parameters and/or data variables and/or model parame-
20 ters from a trained classification system,
 - h) correlating physical parameters and/or data variables and/or model pa-
rameters from the sample with physical parameters and/or data variables
and/or model parameters of the trained system, and
 - 25 i) displaying characterisation class(es) of the sample.
42. The method according to claim 41, wherein the model parameters are latent
variables being weighted averages of the data variables.
- 30 43. The method according to claim 41, wherein the biological sample is selected
from blood, serum, plasma, saliva, urine, cerebrospinal fluid, tears, nasal se-
crete, semen, milk, bile, lymph, sweat and/or faeces.

44. The method according to claim 41, wherein the biological sample is a tissue sample.
45. The method according to claim 41, wherein the tissue sample is a biopsy of tissue selected from muscle, cutis, subcutis, kidney, brain, and liver.
46. The method according to claim 43, wherein the biological sample is urine, blood, milk, or serum.
47. The method according to claim 41, wherein the wavelength of the excitation light is in the range of from 100 nm to 1000 nm, such as from 100 to 800 nm.
48. The method according to claim 41, wherein the wavelength of the excitation light is in the range of from 200 nm to 800 nm, such as from 200 nm to 600 nm.
49. The method according to claim 41, wherein the physical parameter determined is selected from fluorescence intensity, fluorescence lifetime, phosphorescence intensity, phosphorescence lifetime, polarisation, polarisation lifetime, anisotropy, anisotropy lifetime, phase-resolved emission, circularly polarised fluorescence, fluorescence-detected circular dichroism, and any time dependence of the two last mentioned parameters.
50. The method according to claim 41, wherein the spectral distribution of light emitted ranging from 200 nm to 800 nm is generated
51. The method according to claim 41, wherein the data handling of step d) is selected from a one-way matrix of spectral information, a two-way matrix of spectral information, a three-way matrix of spectral information, a four-way matrix of spectral information and, a five-way or higher order matrix of spectral information.
52. The method according to claim 41, wherein other variable(s) is included as data variables.

53. The method according to claim 52, wherein the other variable(s) is selected from a pH value of the sample, concentration of various electrolytes in the sample, concentration of any other relevant compound in the sample, temperature, chemical parameters or any other physical property of the sample.
- 5
54. The method according to claim 41, wherein other variable(s) related to the animal, including a human being, is included as data variables
55. The method according to claim 54, wherein the other variable(s) is selected from any parameter relating to the bodily or mental condition, hair colour, skin colour, age, sex, geographic origin, affiliation, hereditary background, stress level, medical diagnosis, subjective evaluations or clinical parameters.
- 10
56. The method according to claim 41, wherein the sample is pre-treated before subjecting the sample to step b).
- 15
57. The method according to claim 56, wherein the pre-treatment comprises adjustment of pH of the sample to a predetermined value.
58. The method according to claim 41, wherein the trained classification system is a diagnostic heart disease classification system.
- 20
59. The method according to claim 41, wherein the trained classification system is a diagnostic abuse classification system related to abuse of medicine or narcotics.

Fig. 1

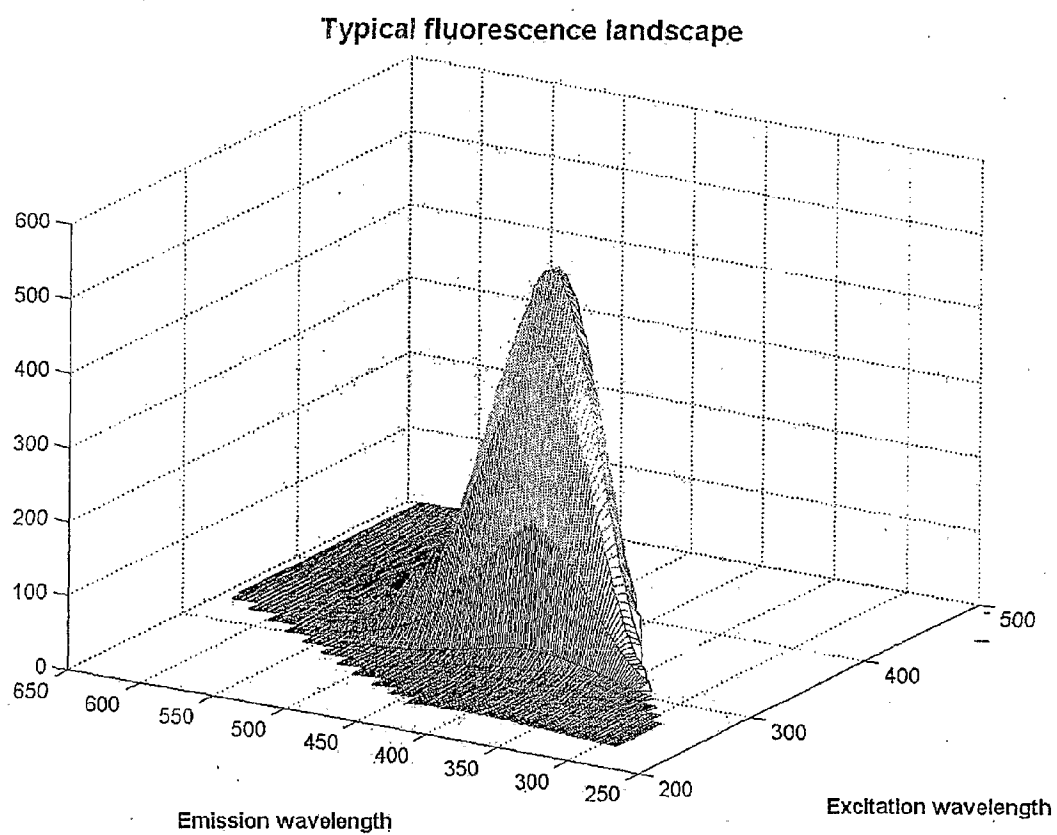


Fig. 2

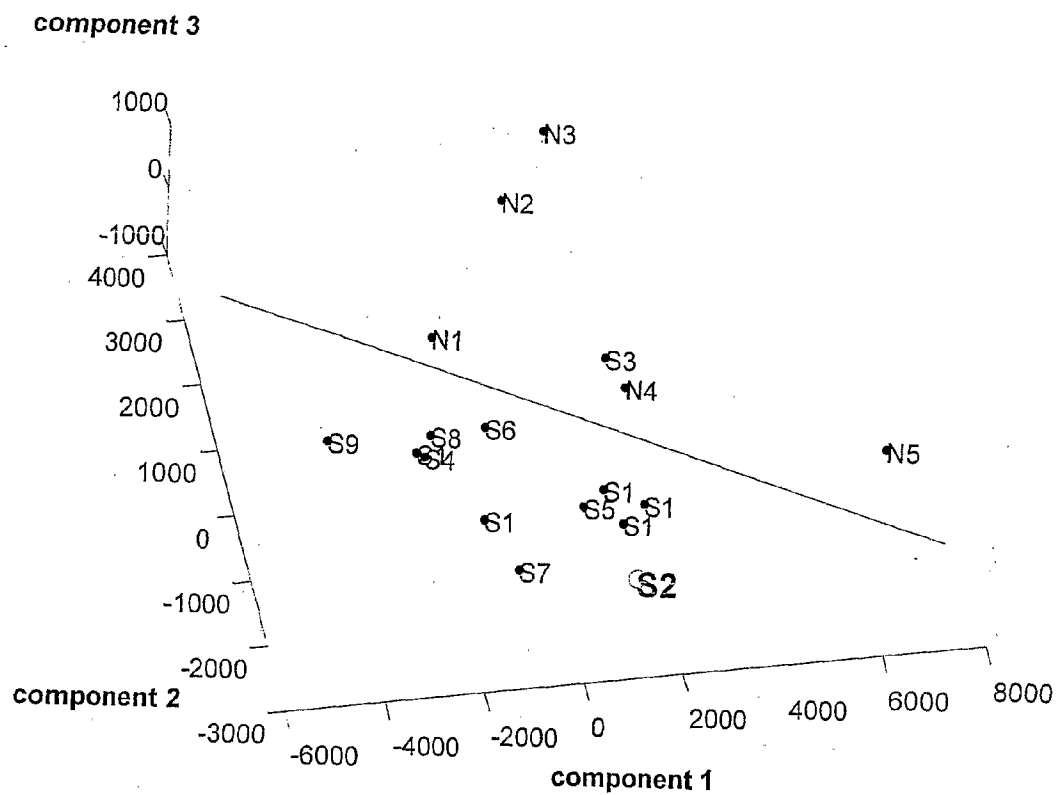
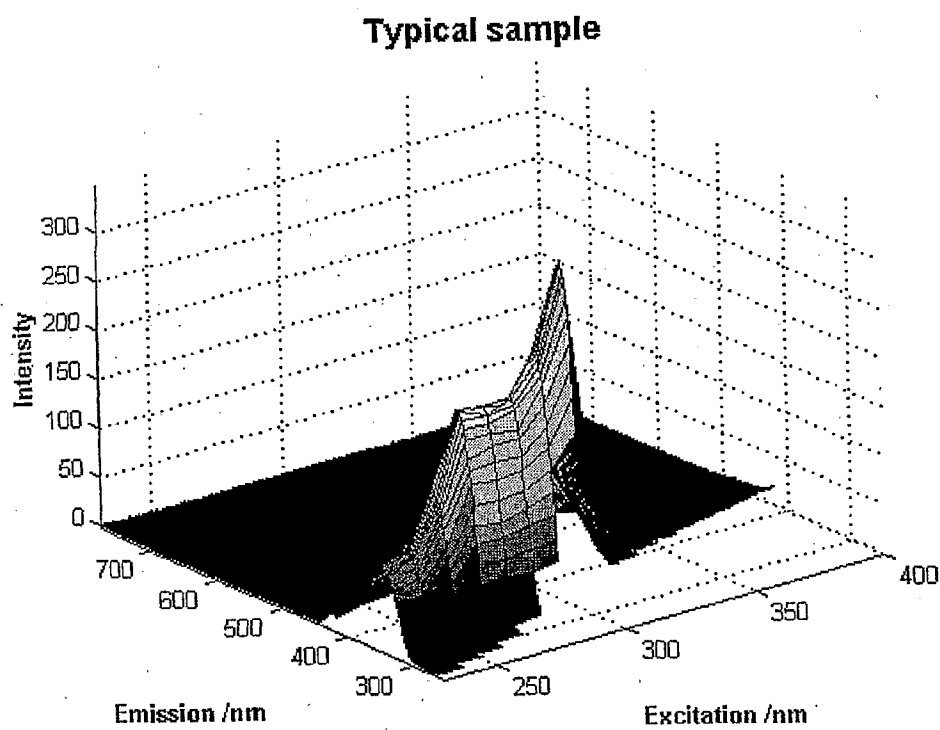


Fig. 3



4/36

Fig. 4

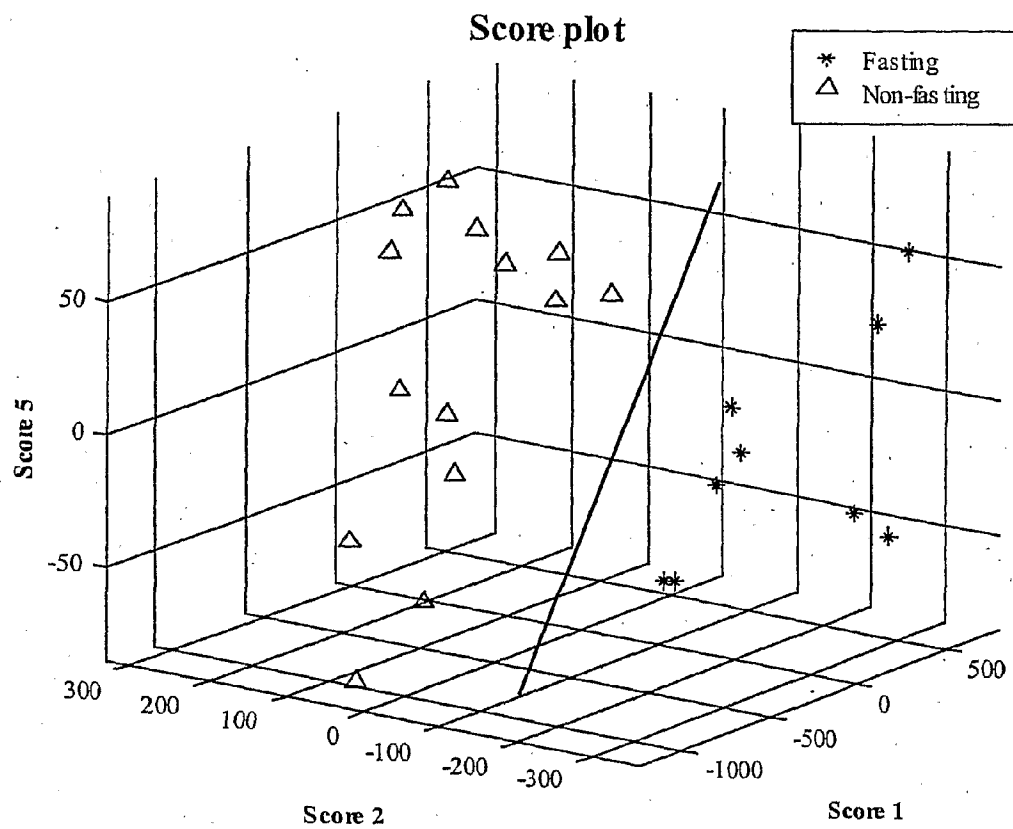
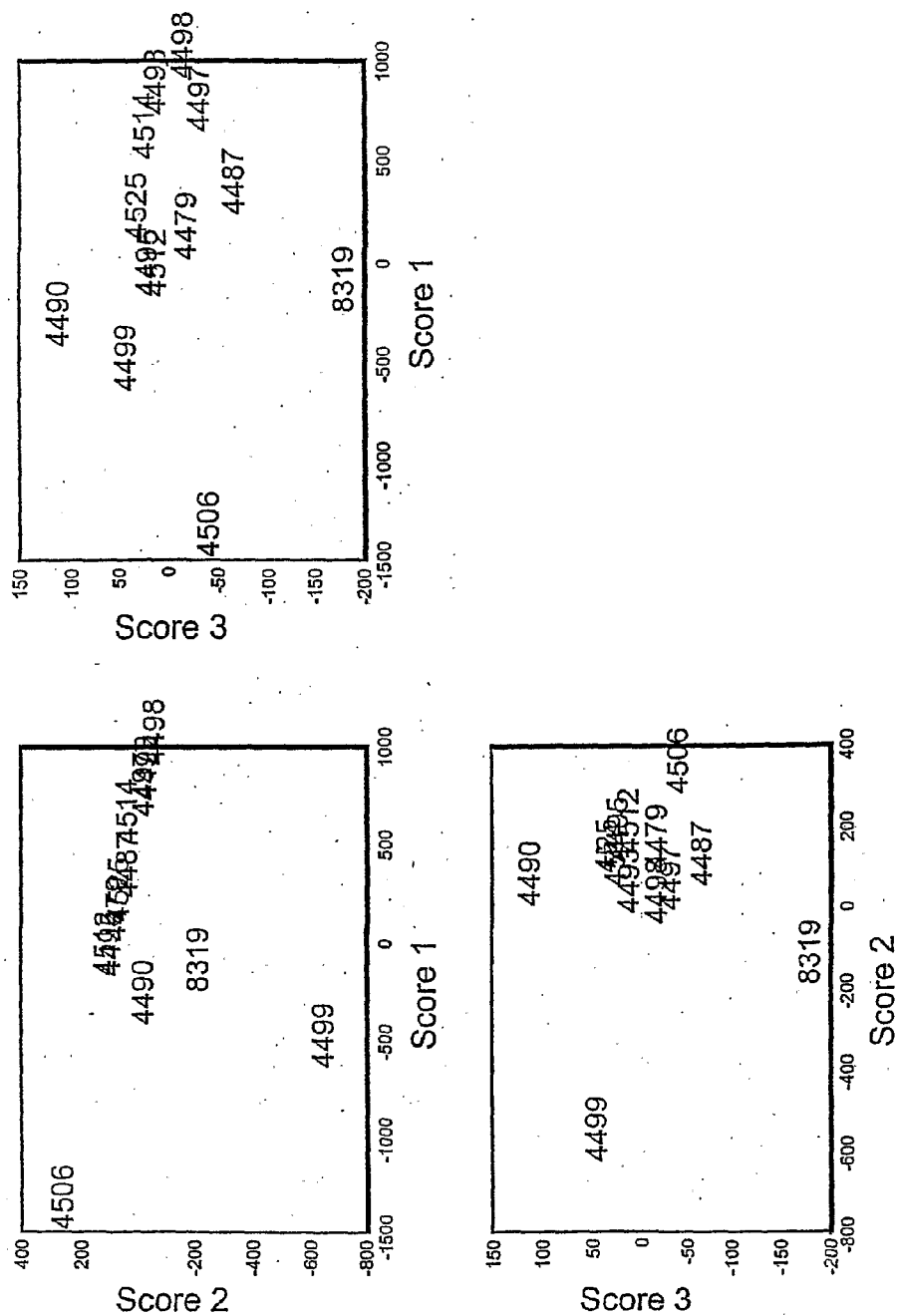
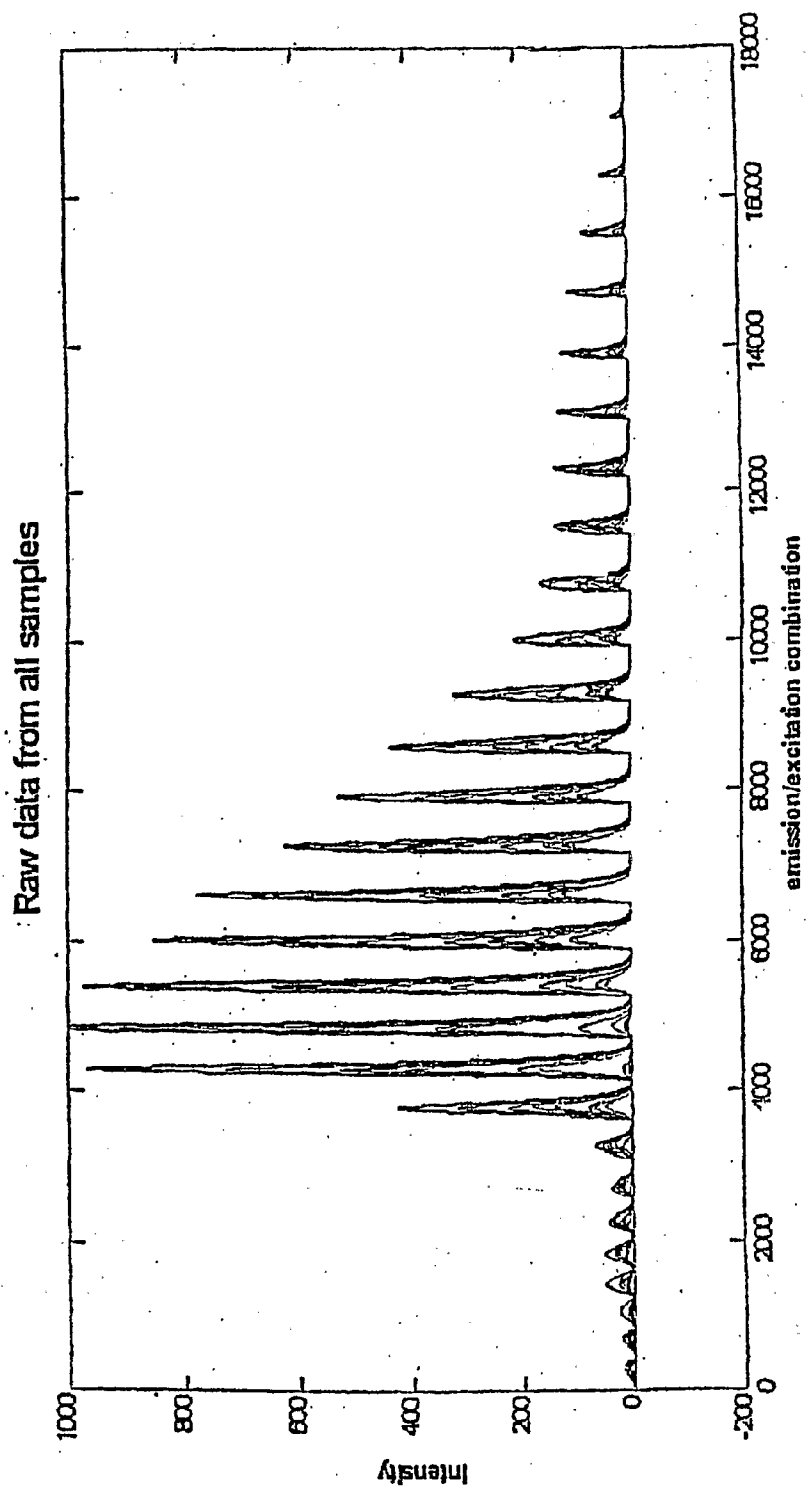


Fig. 5



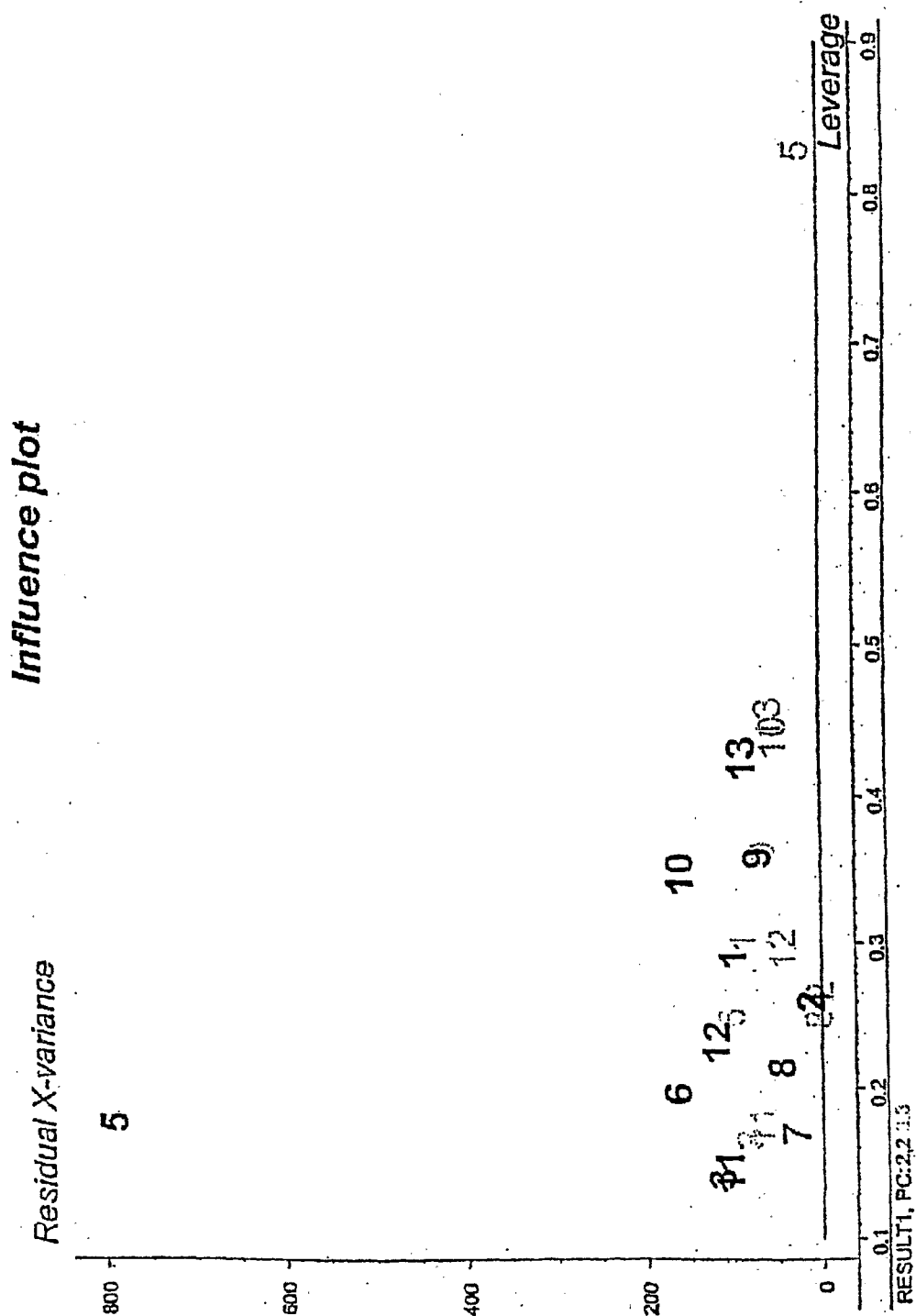
6/36

Fig. 6



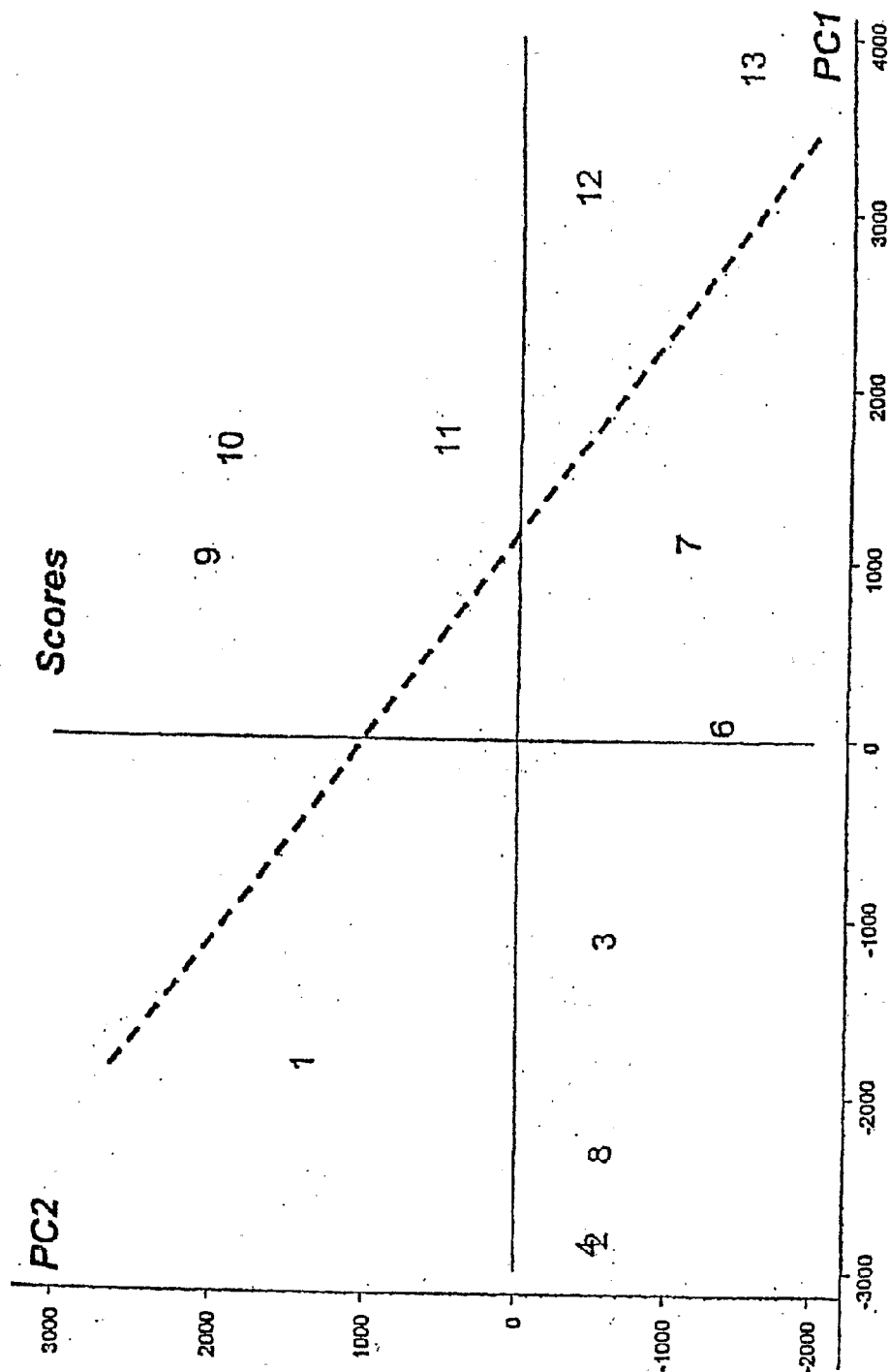
7/36

Fig. 7



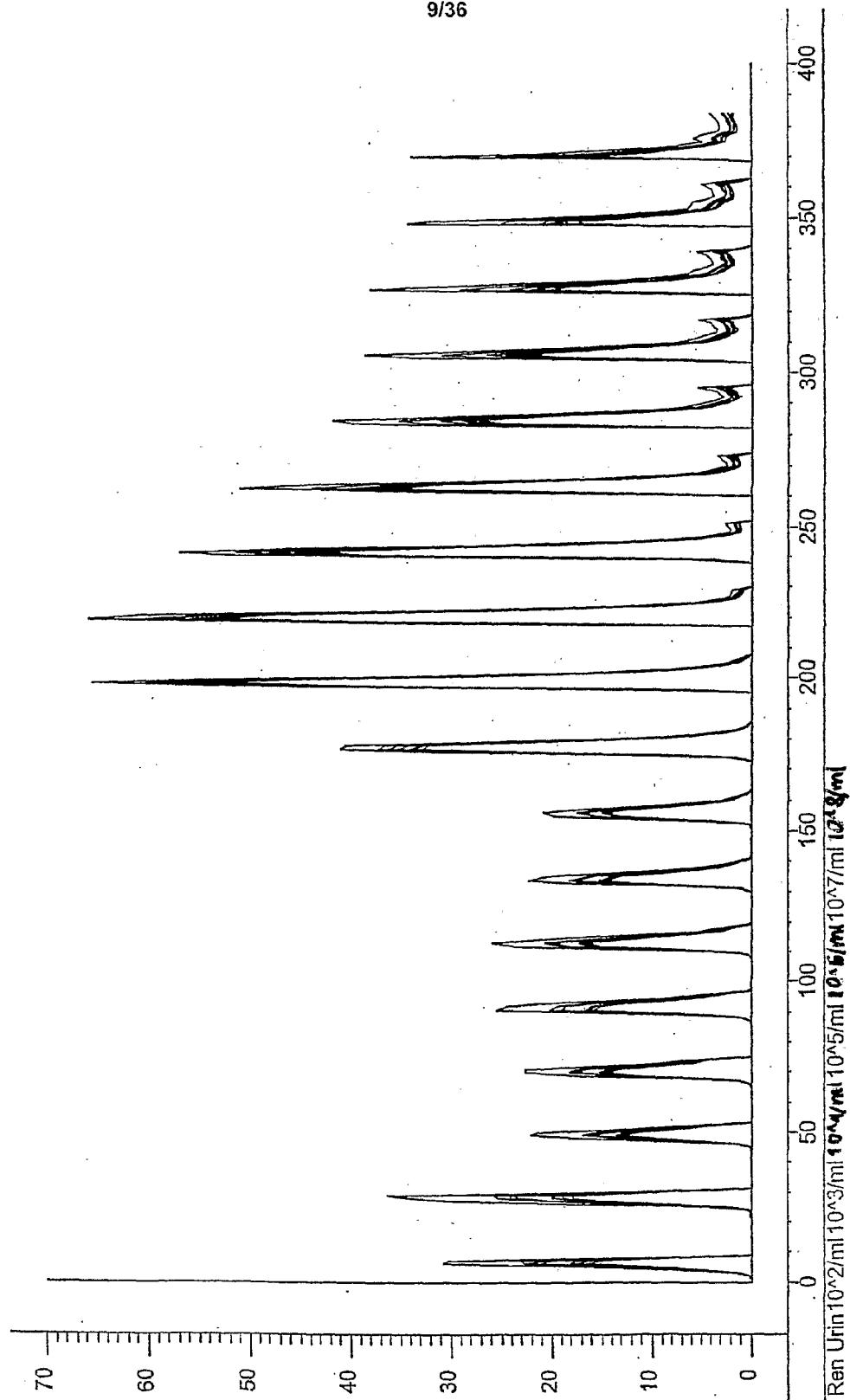
8/36

Fig. 8



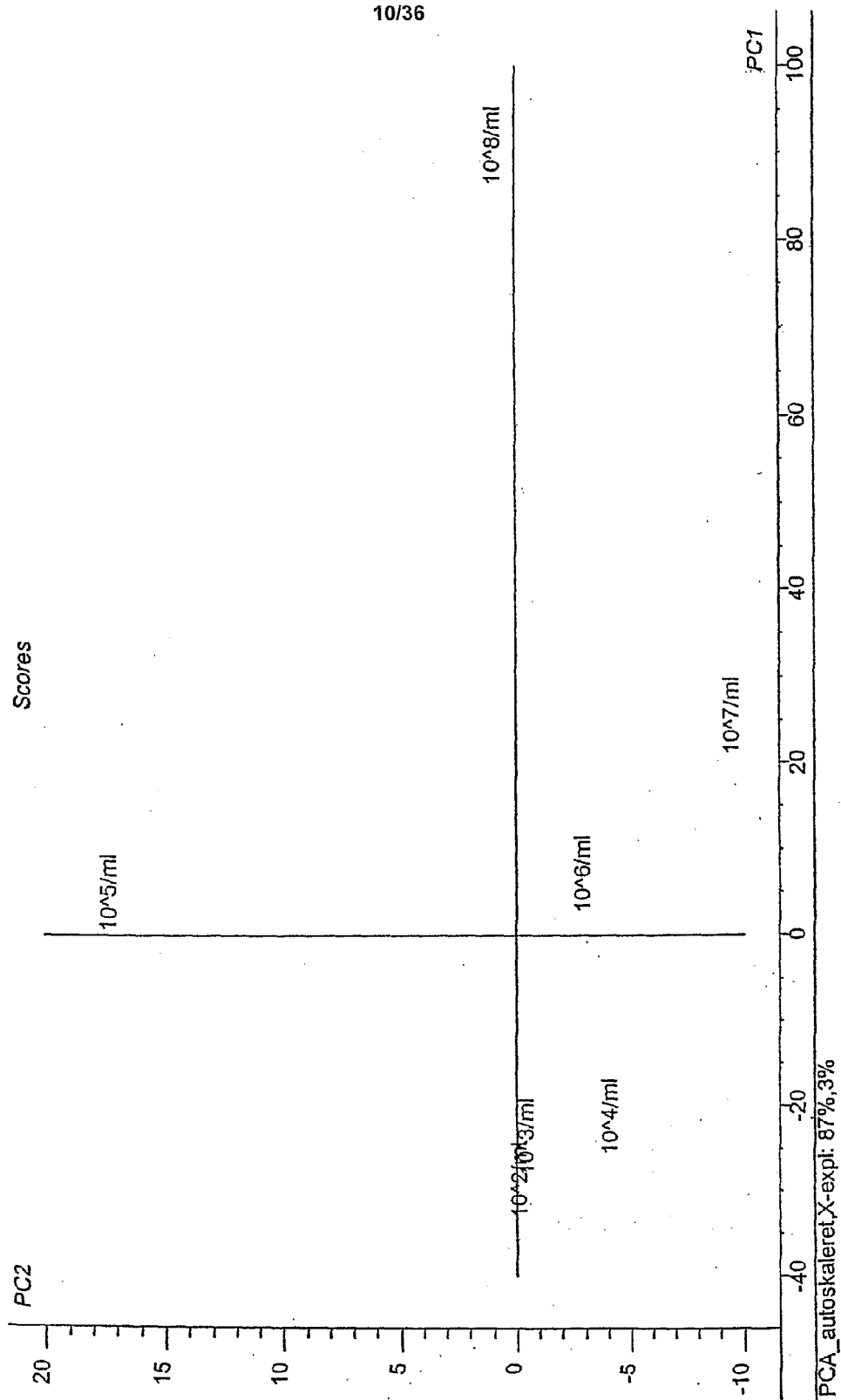
9/36

Fig. 9



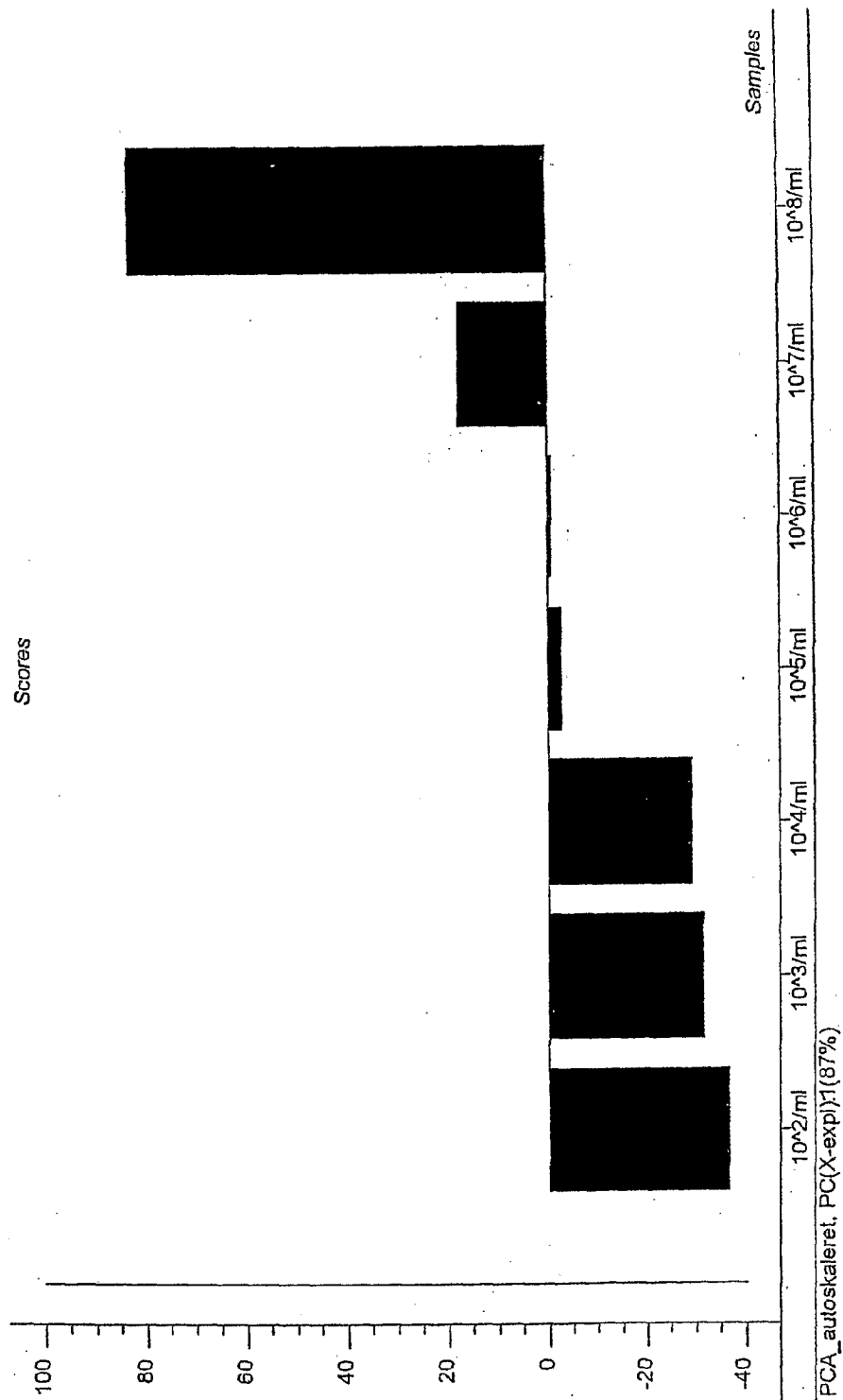
10/36

Fig. 10A



11/36

Fig. 10B



12/36

Fig. 11A

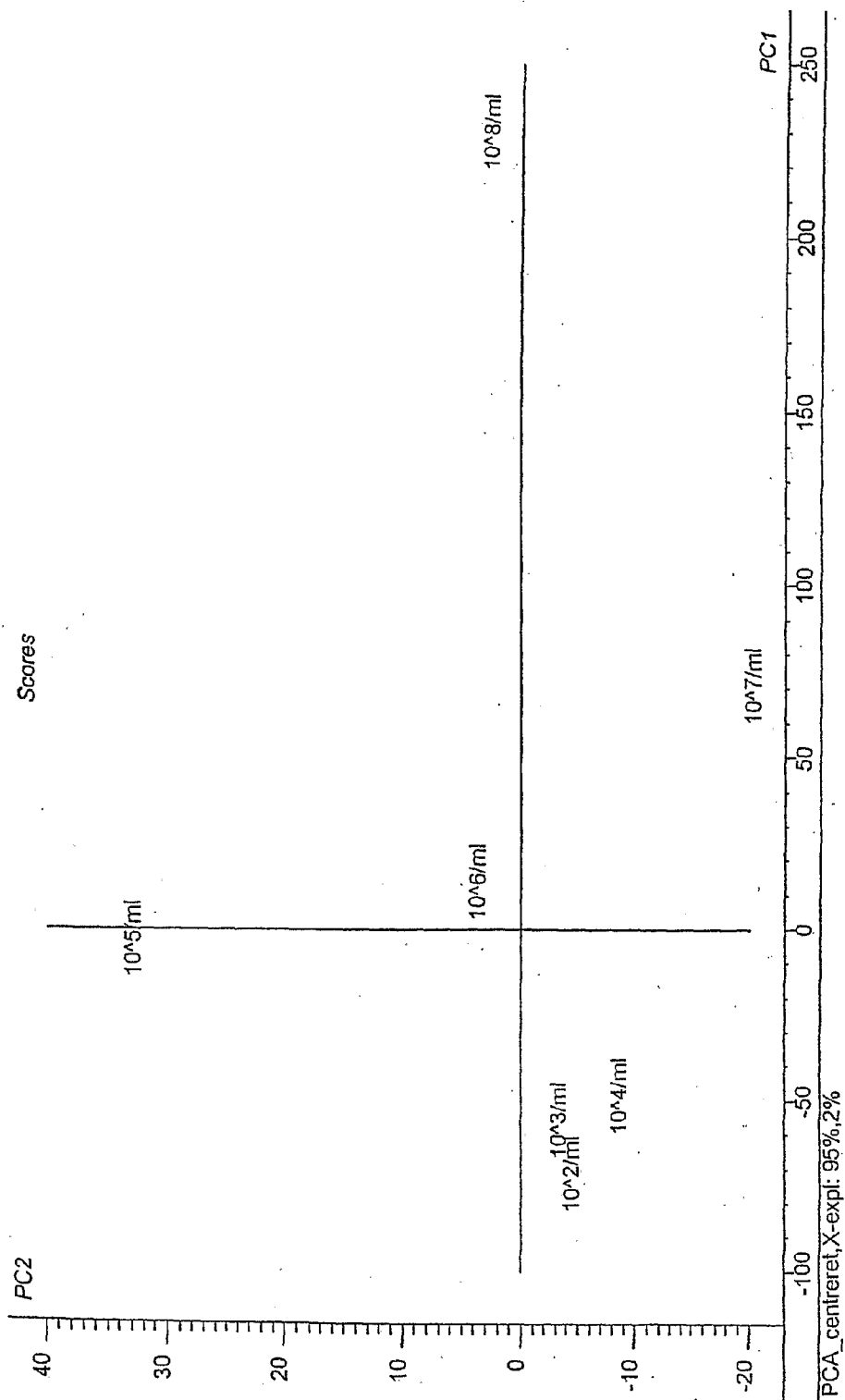


Fig. 11B

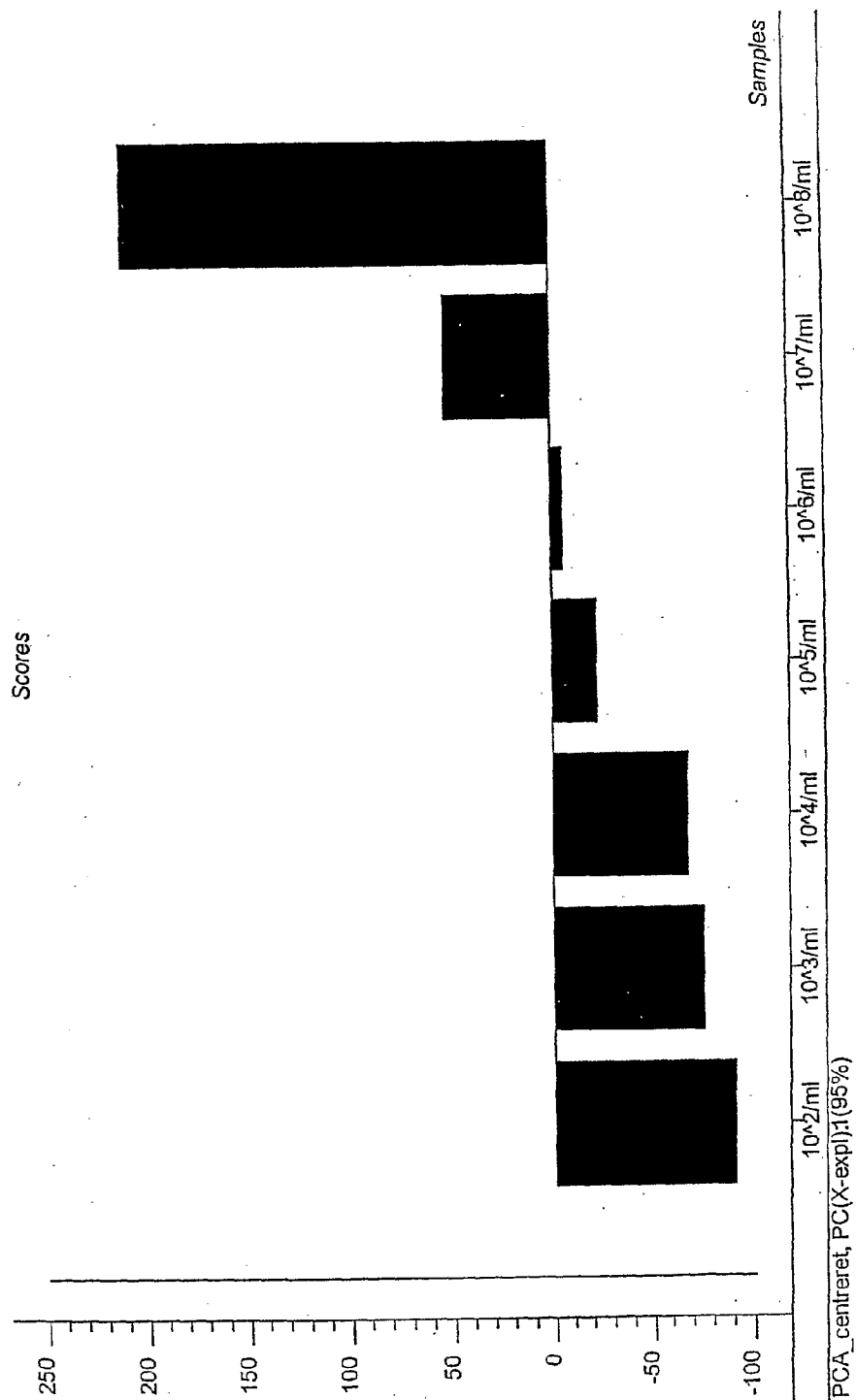
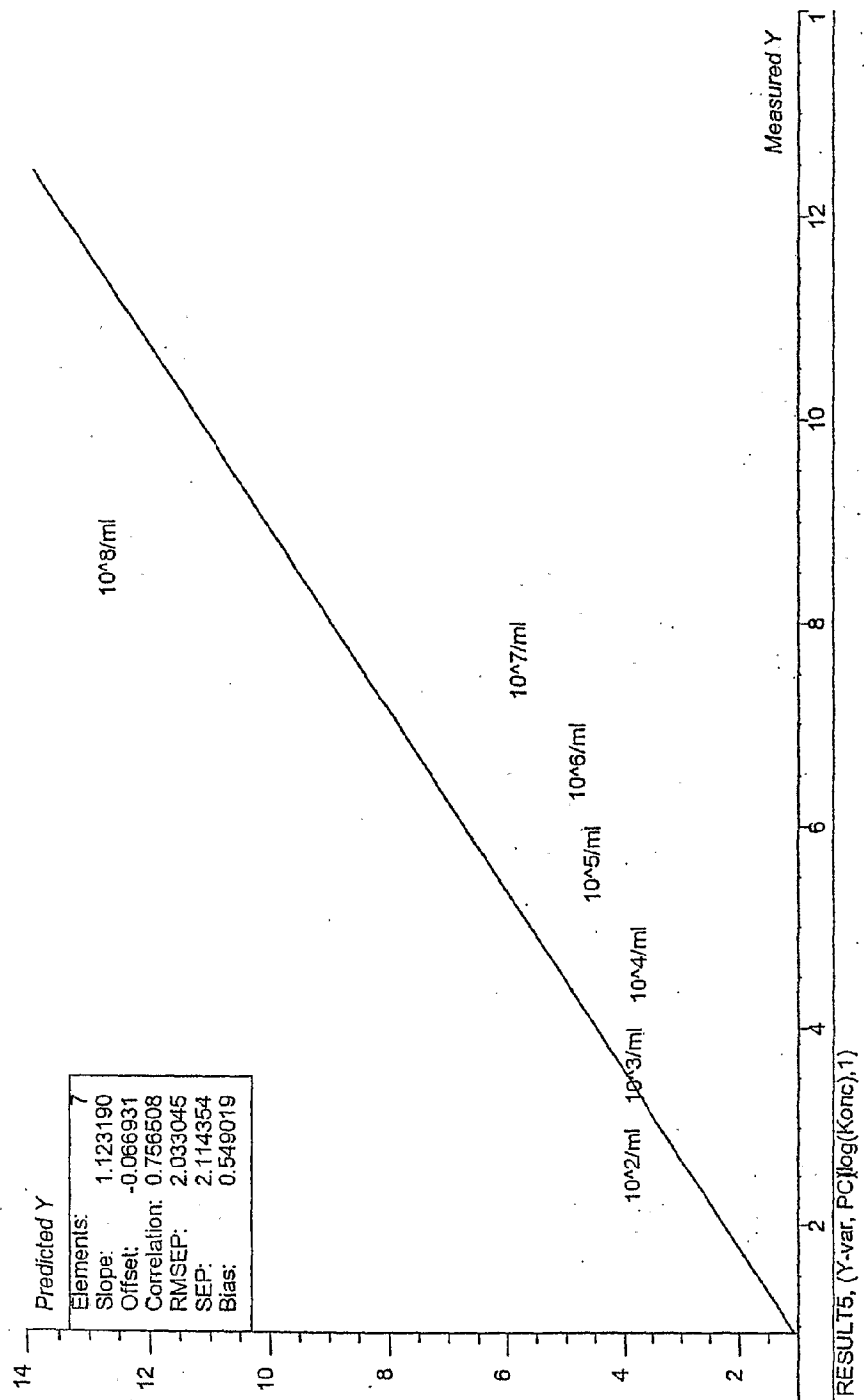


Fig. 12A



15/36

Fig. 12B

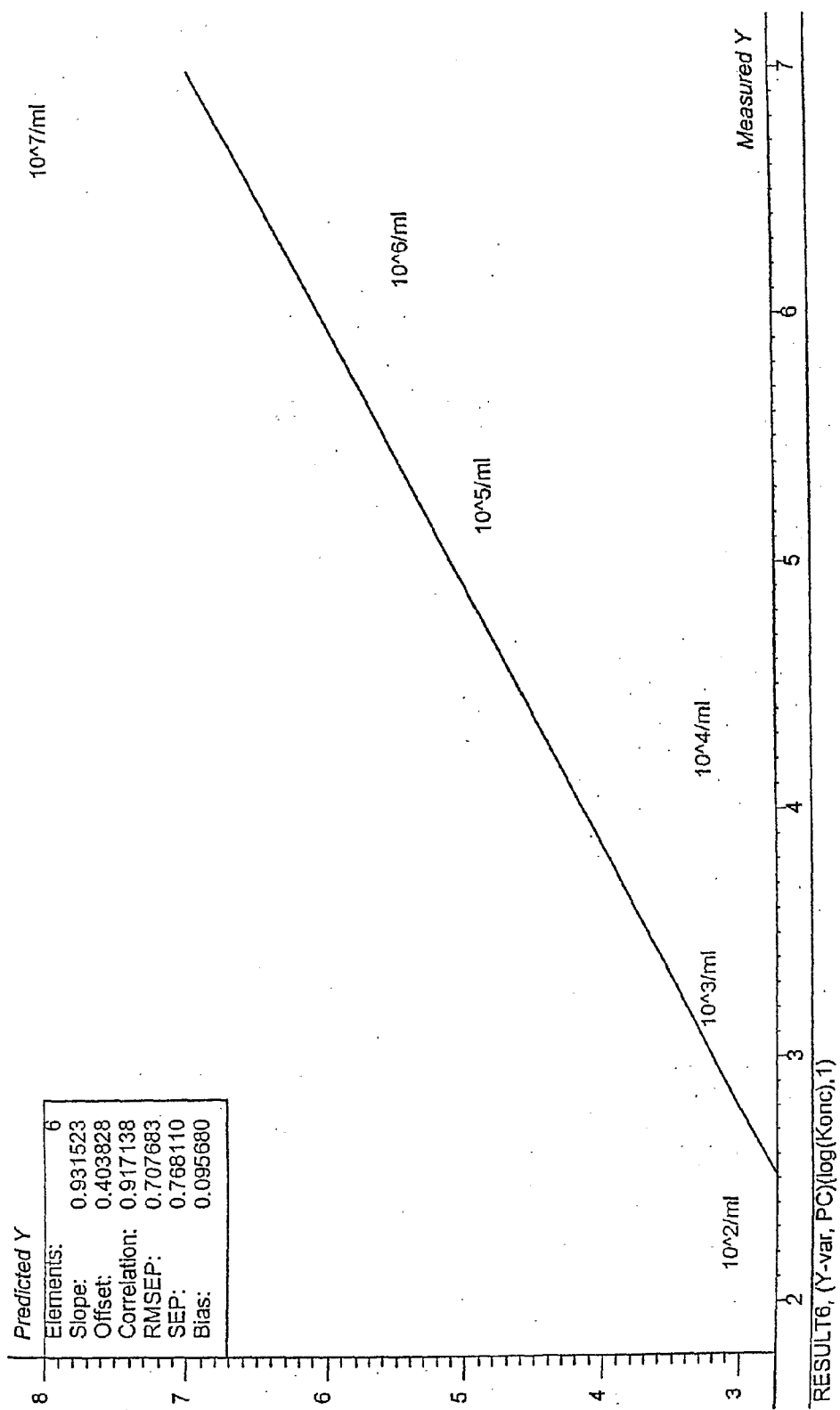


Fig. 13

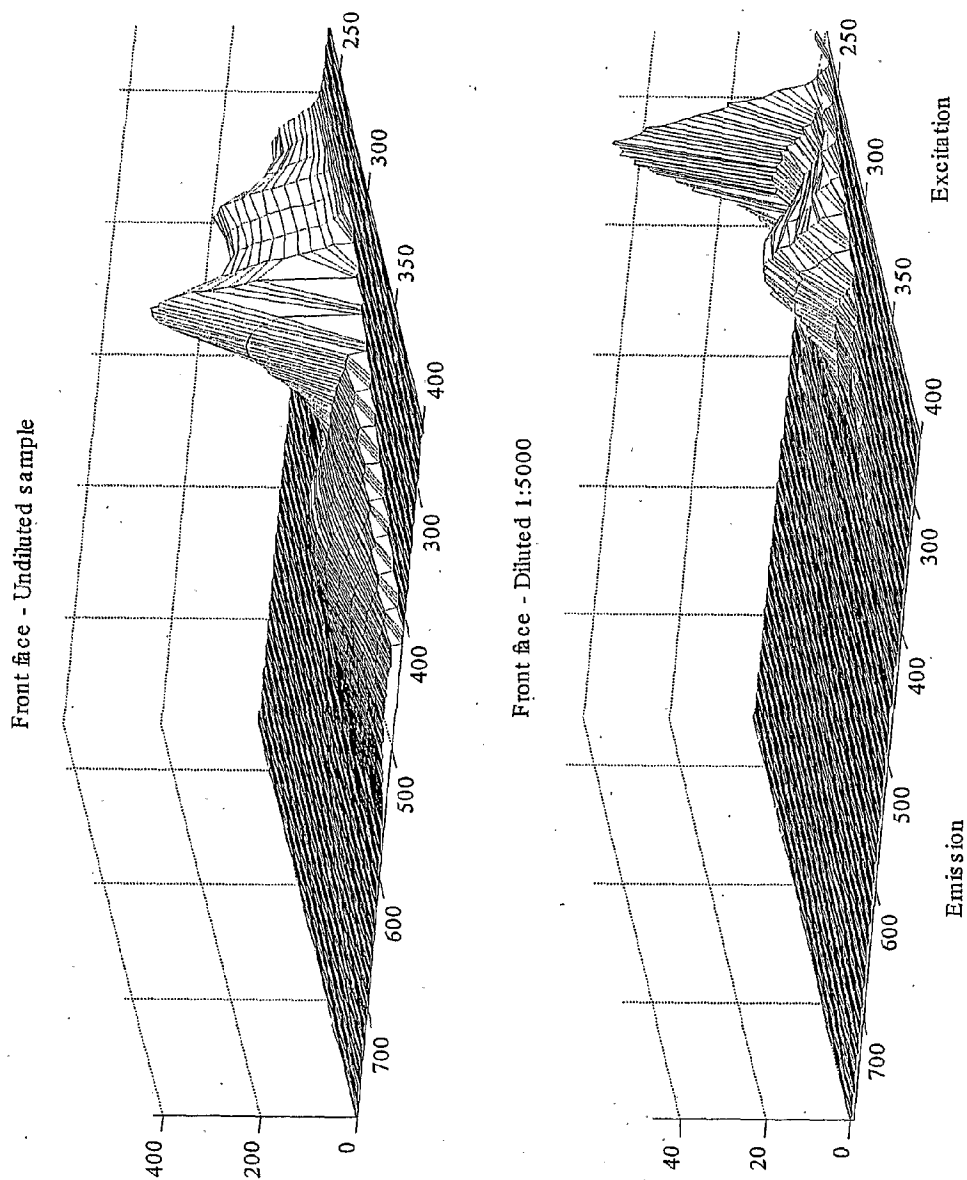


Fig. 14A

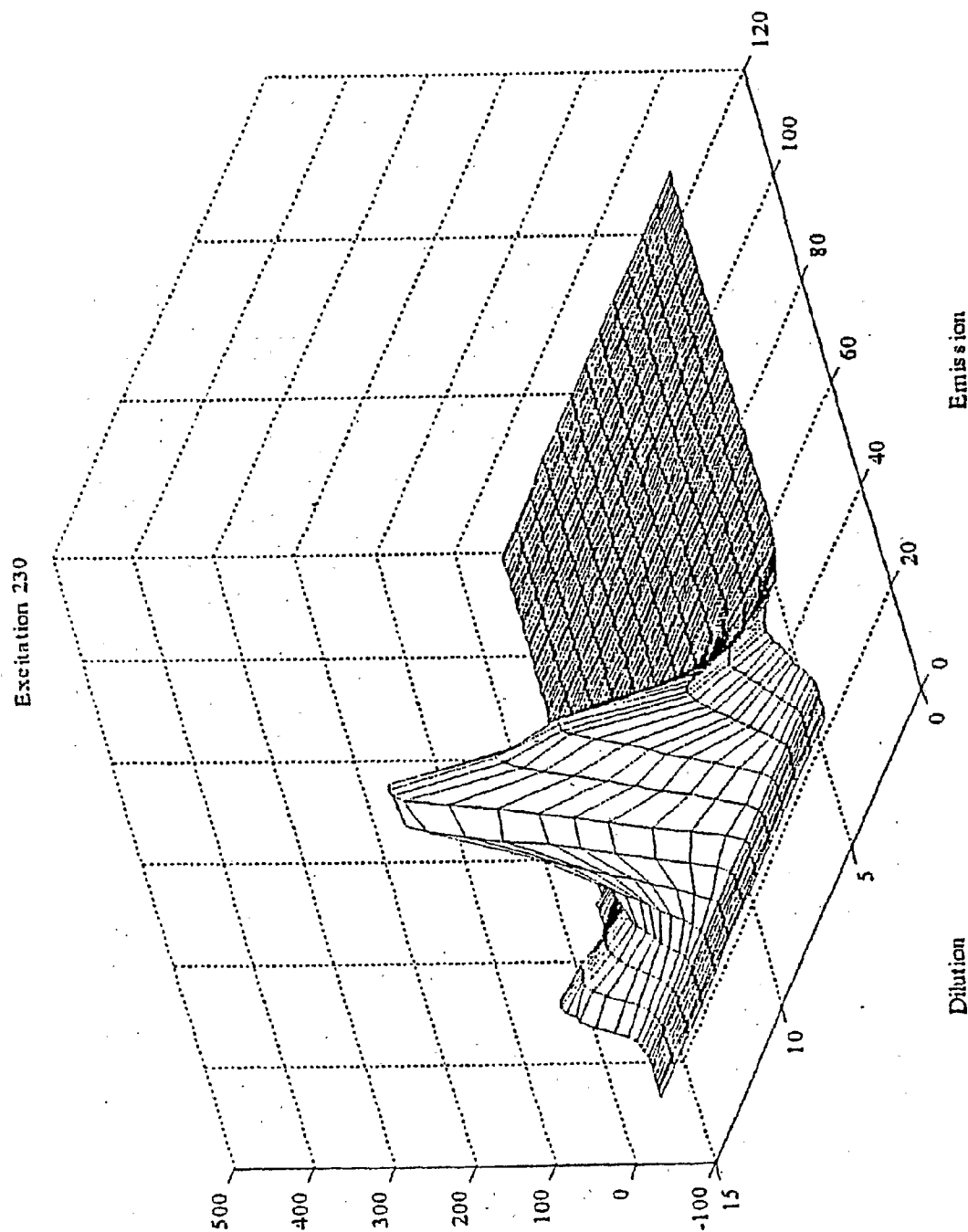


Fig. 14B

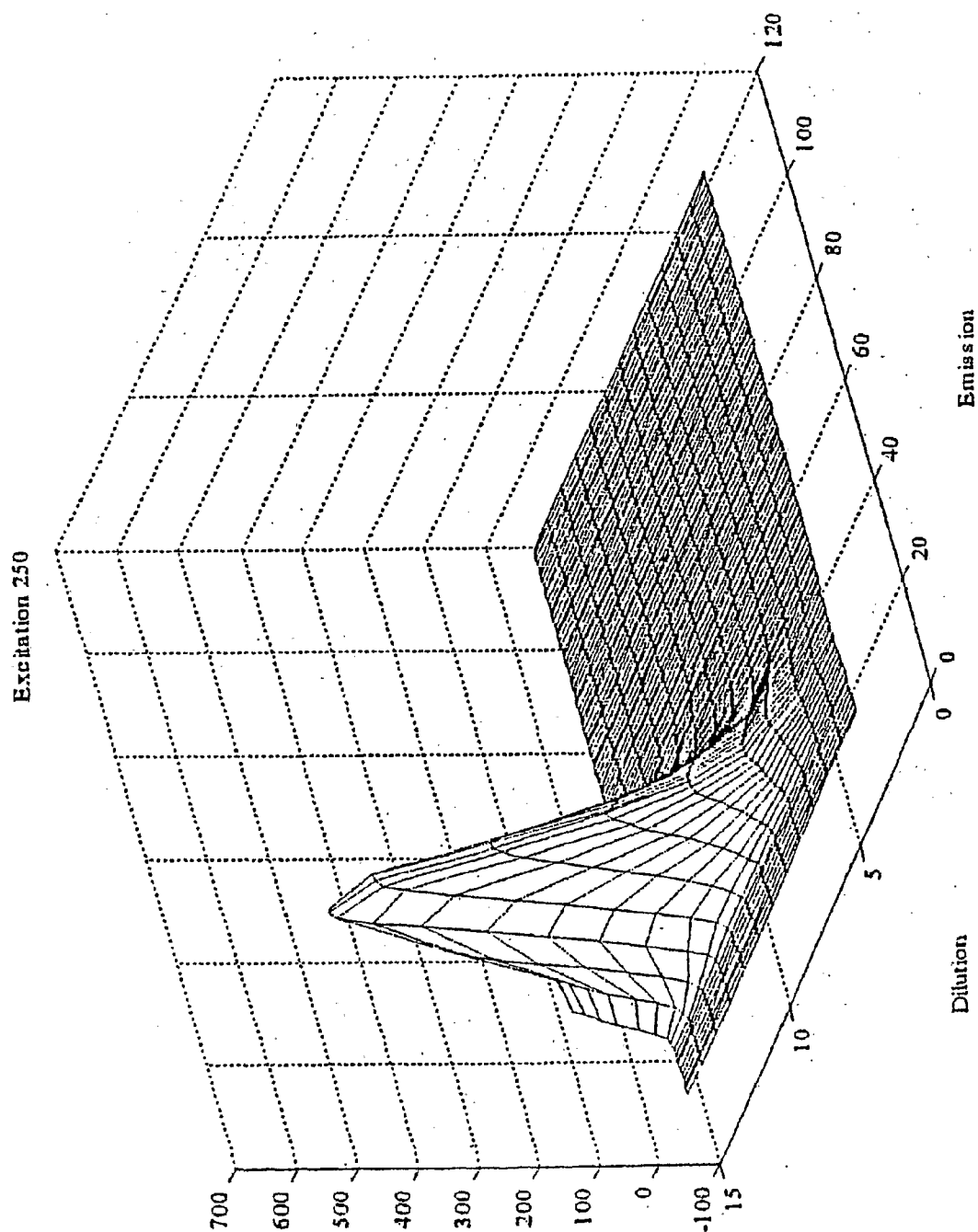


Fig. 15A

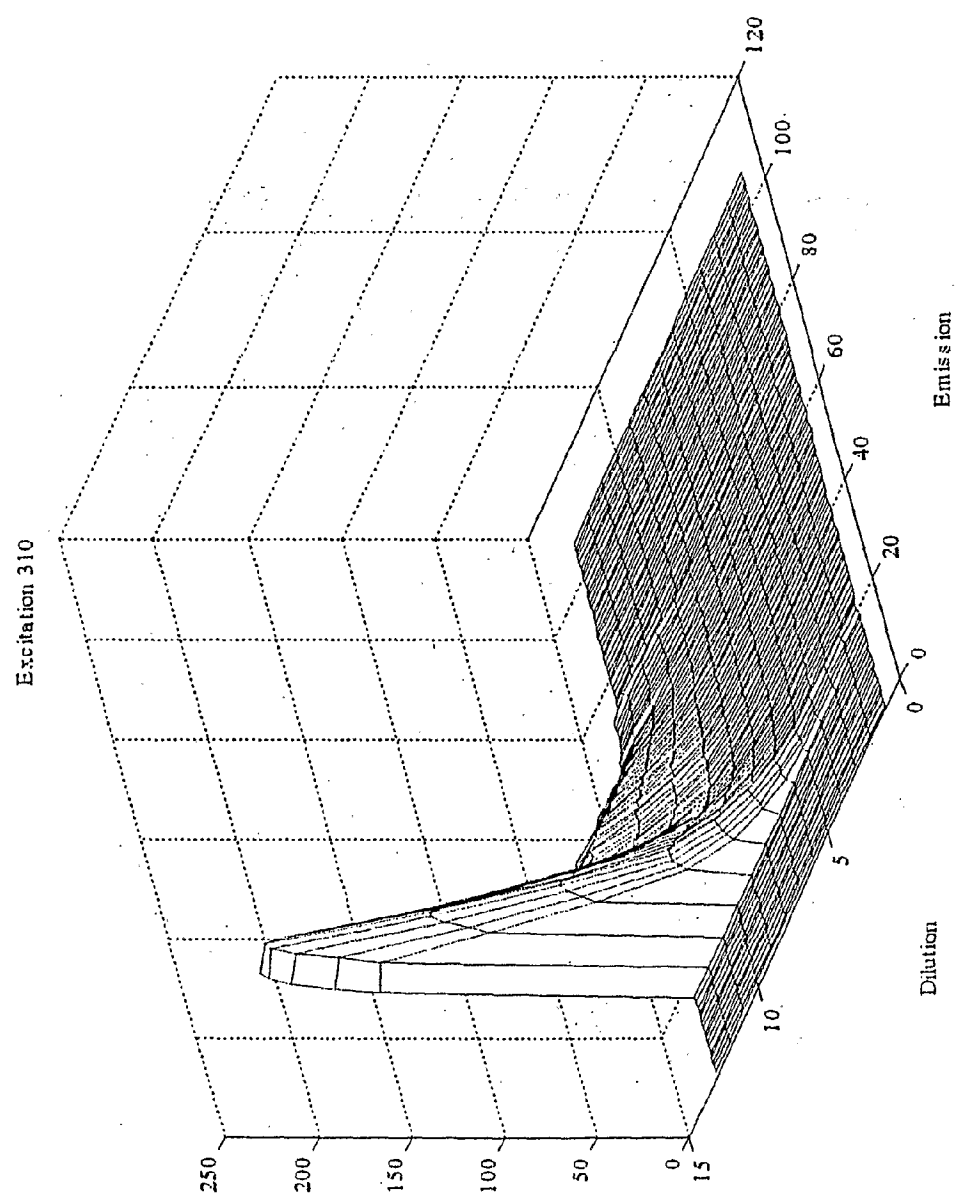


Fig. 15B

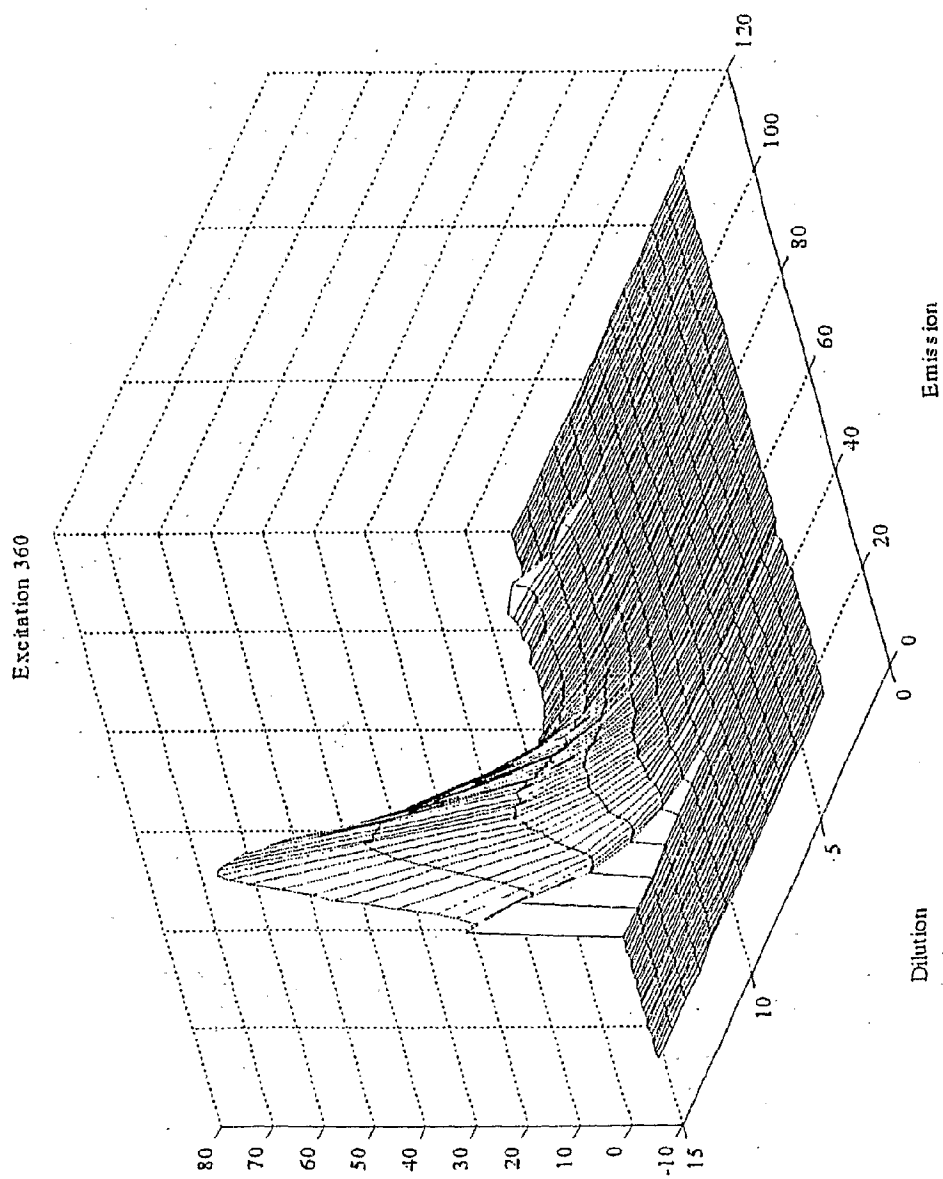
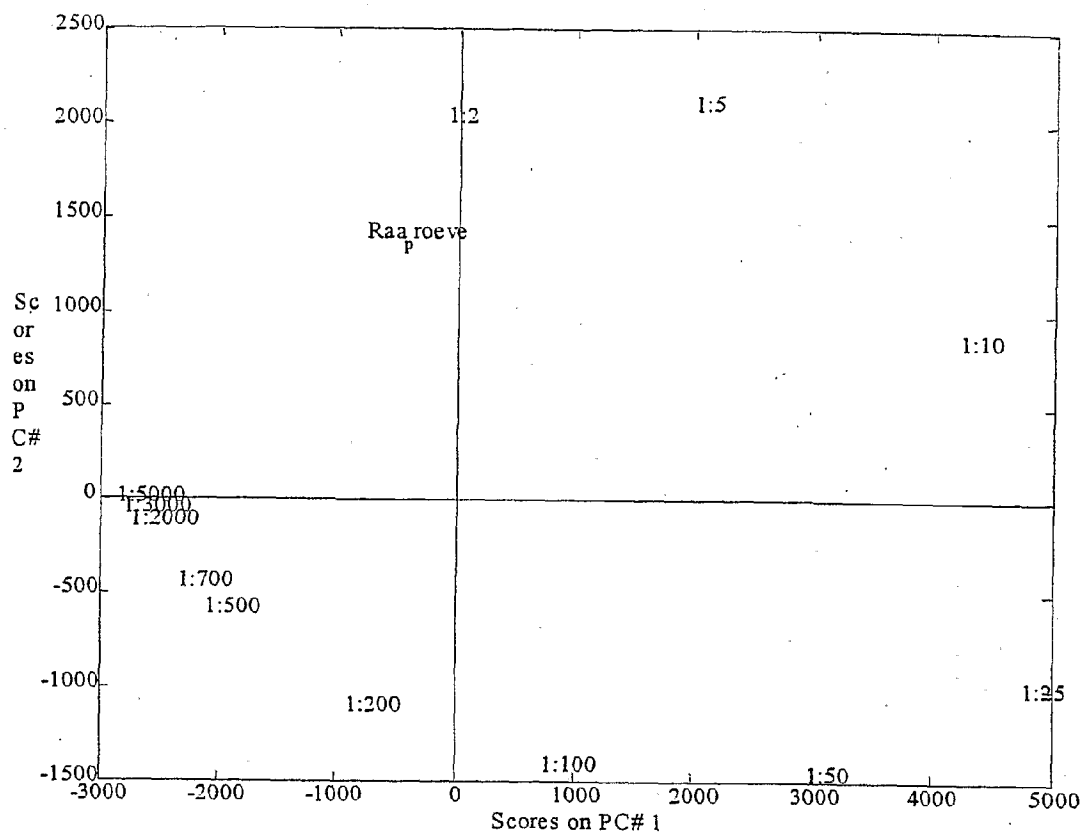
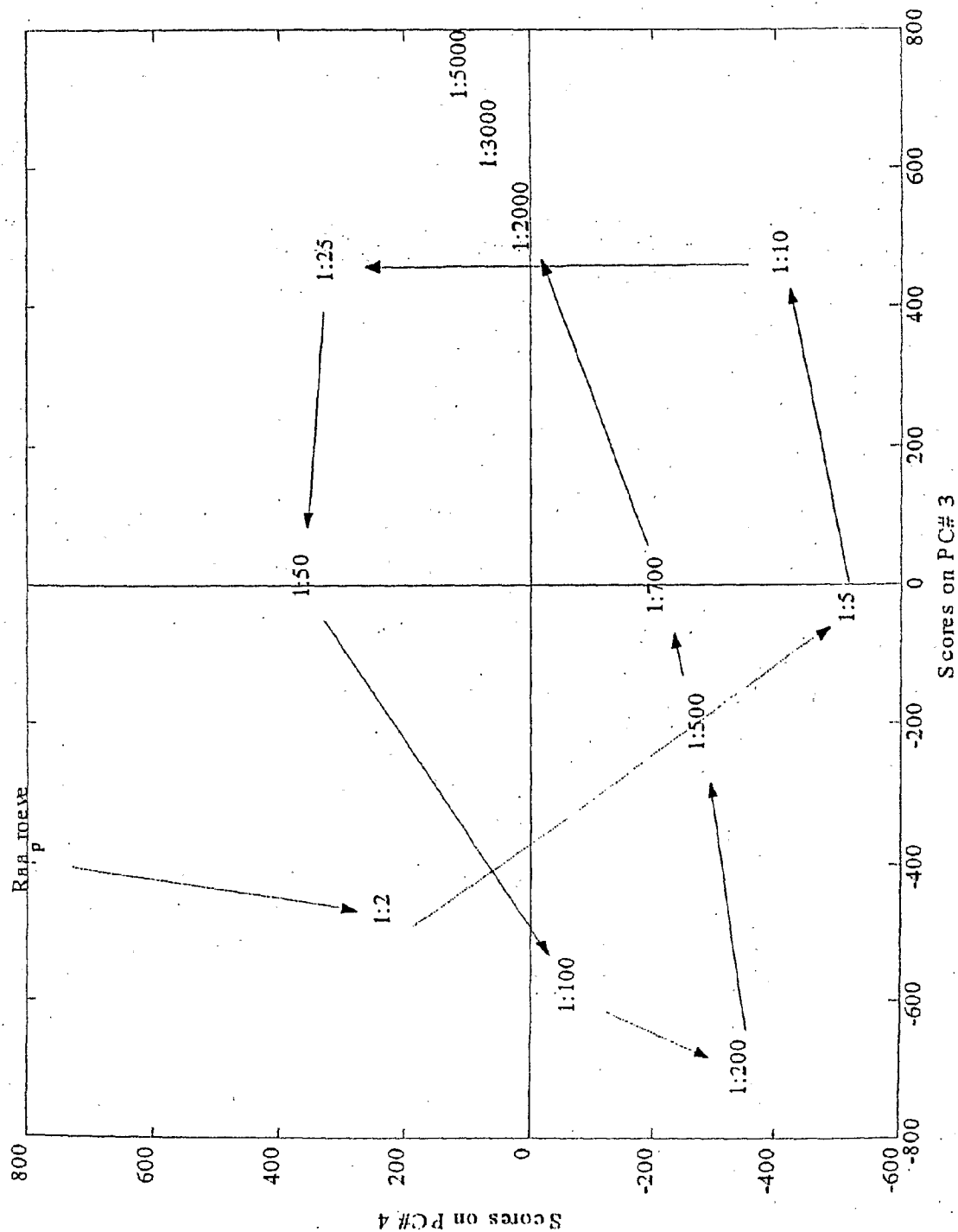


Fig. 16A



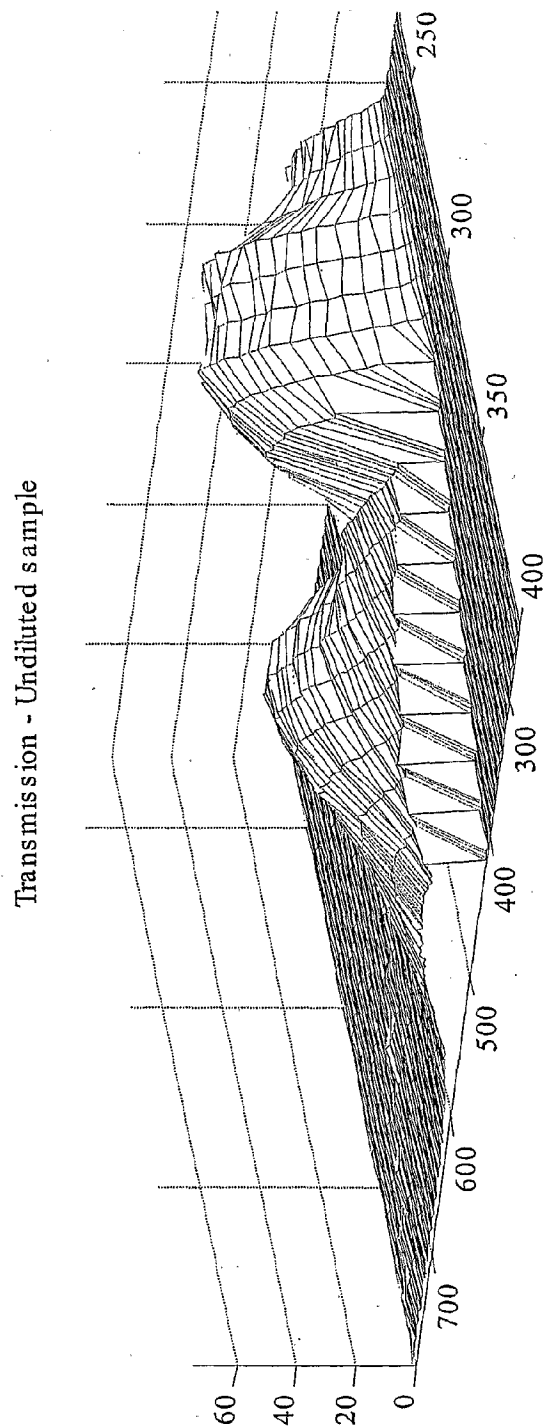
22/36

Fig. 16B



23/36

Fig. 17A



24/36

Fig. 17B

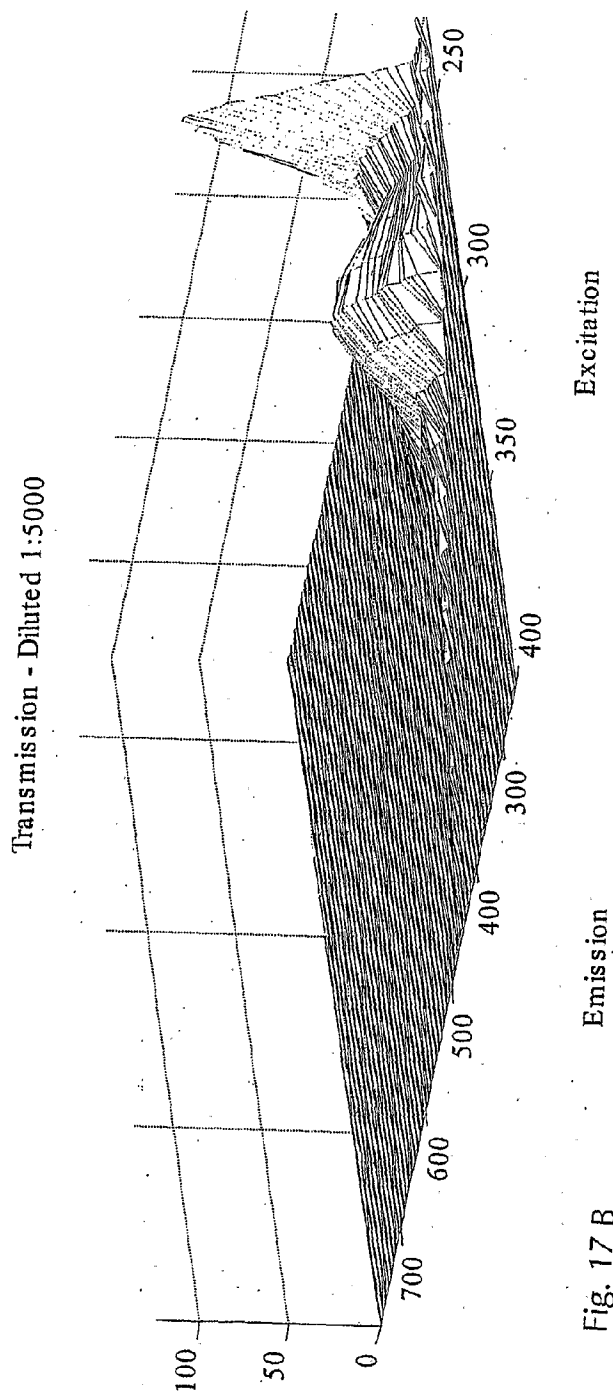


Fig. 17 B

Fig. 18A

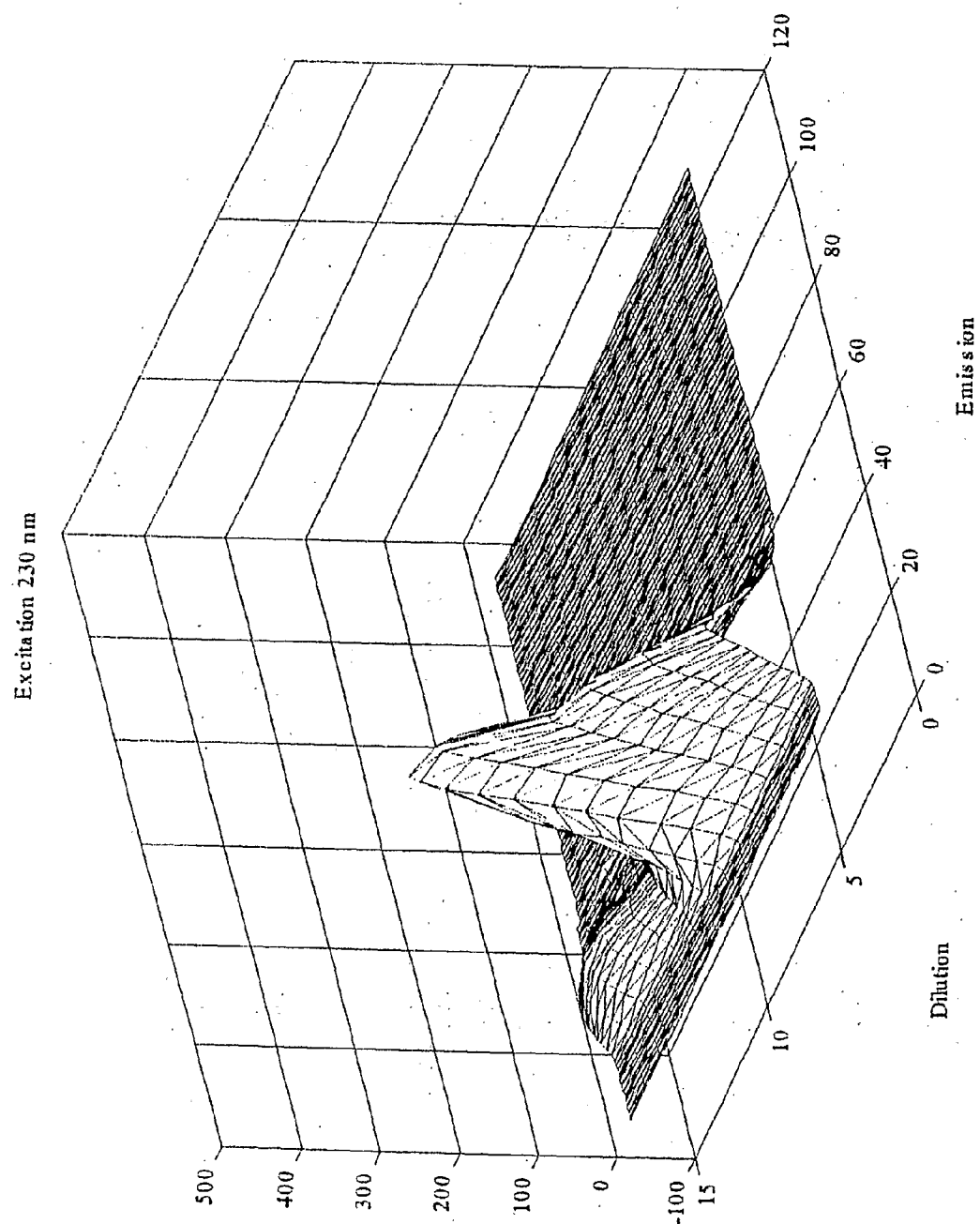


Fig. 18B

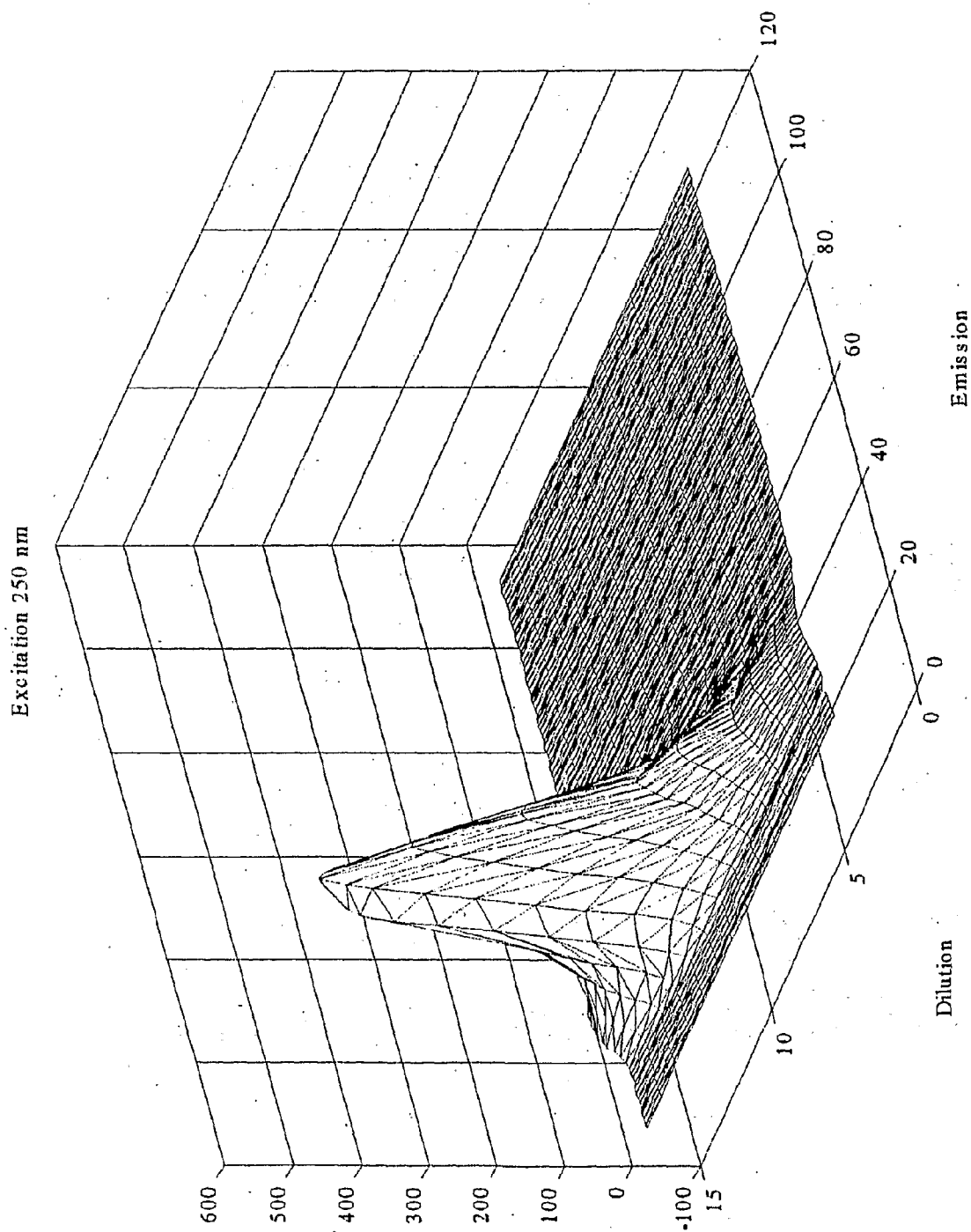


Fig. 19A

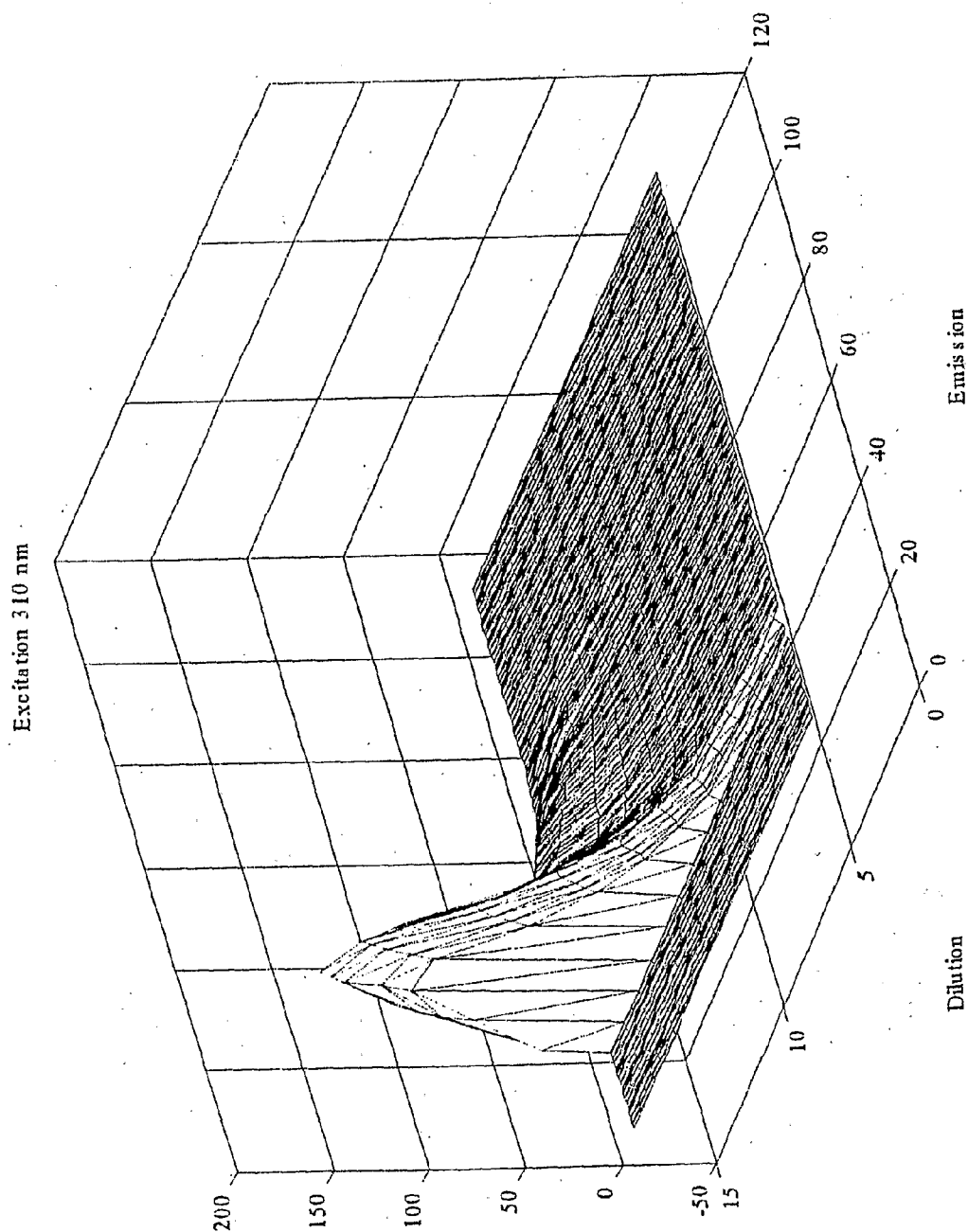
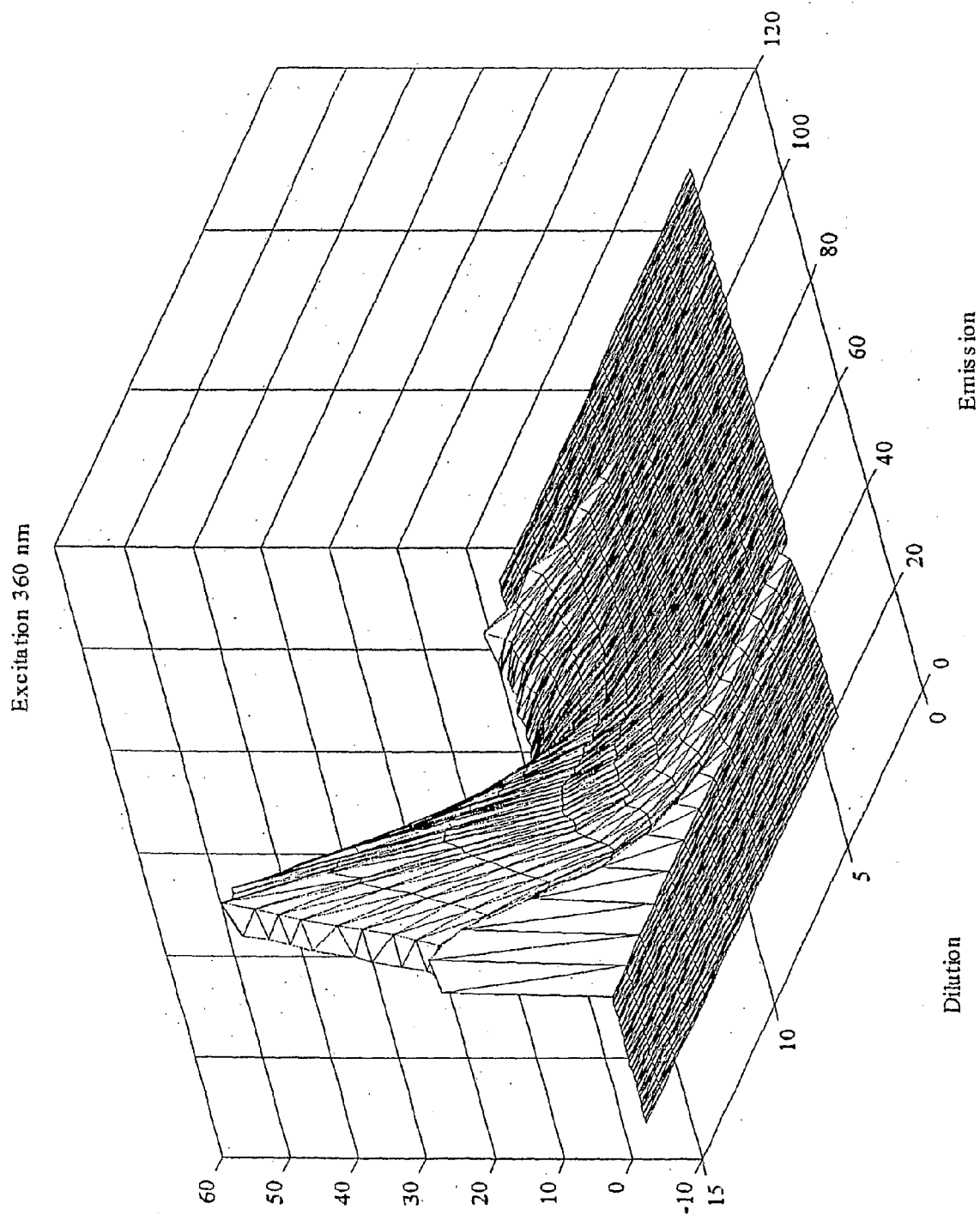


Fig. 19B



29/36

Fig. 20A

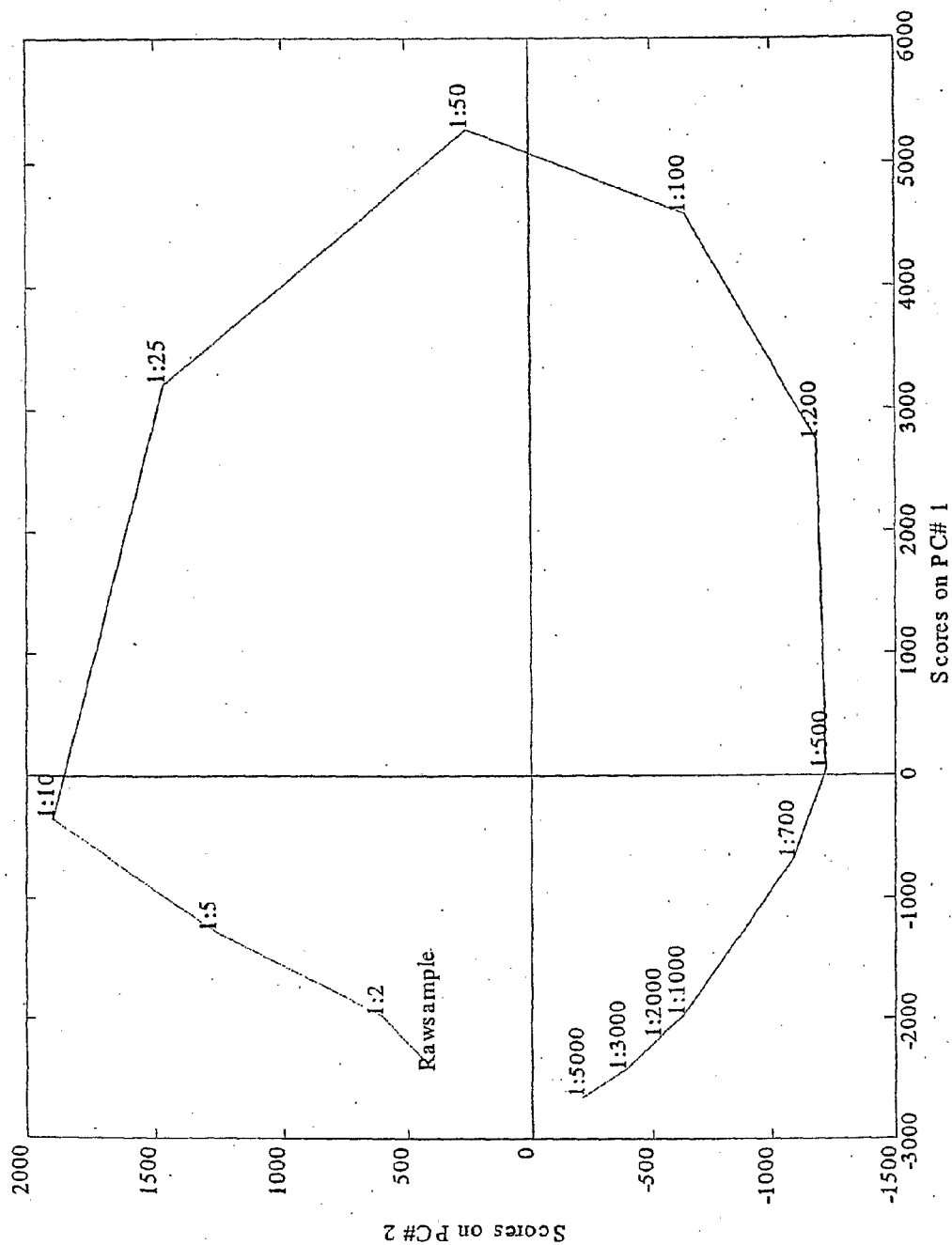


Fig. 20B

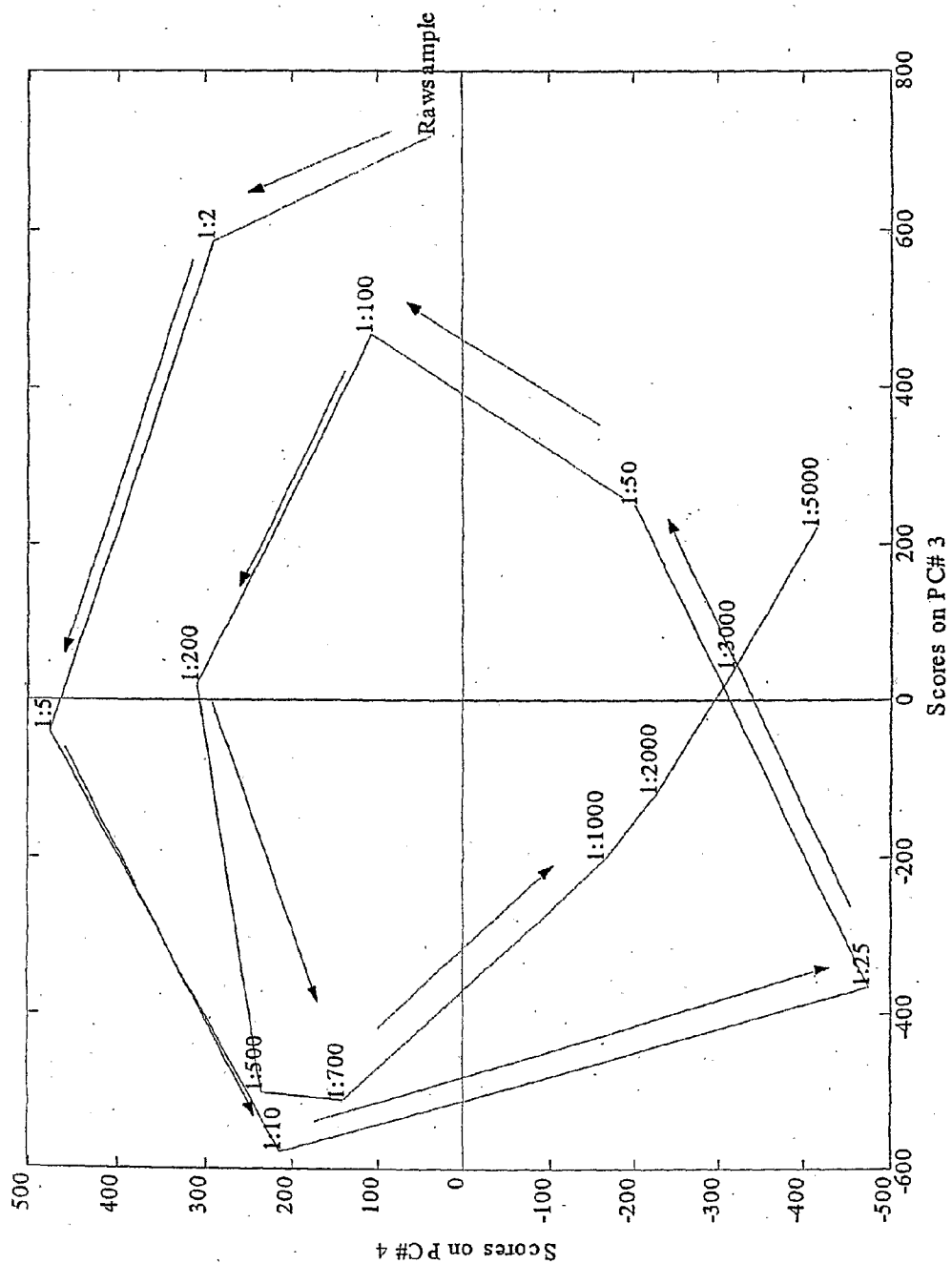


Fig. 21A

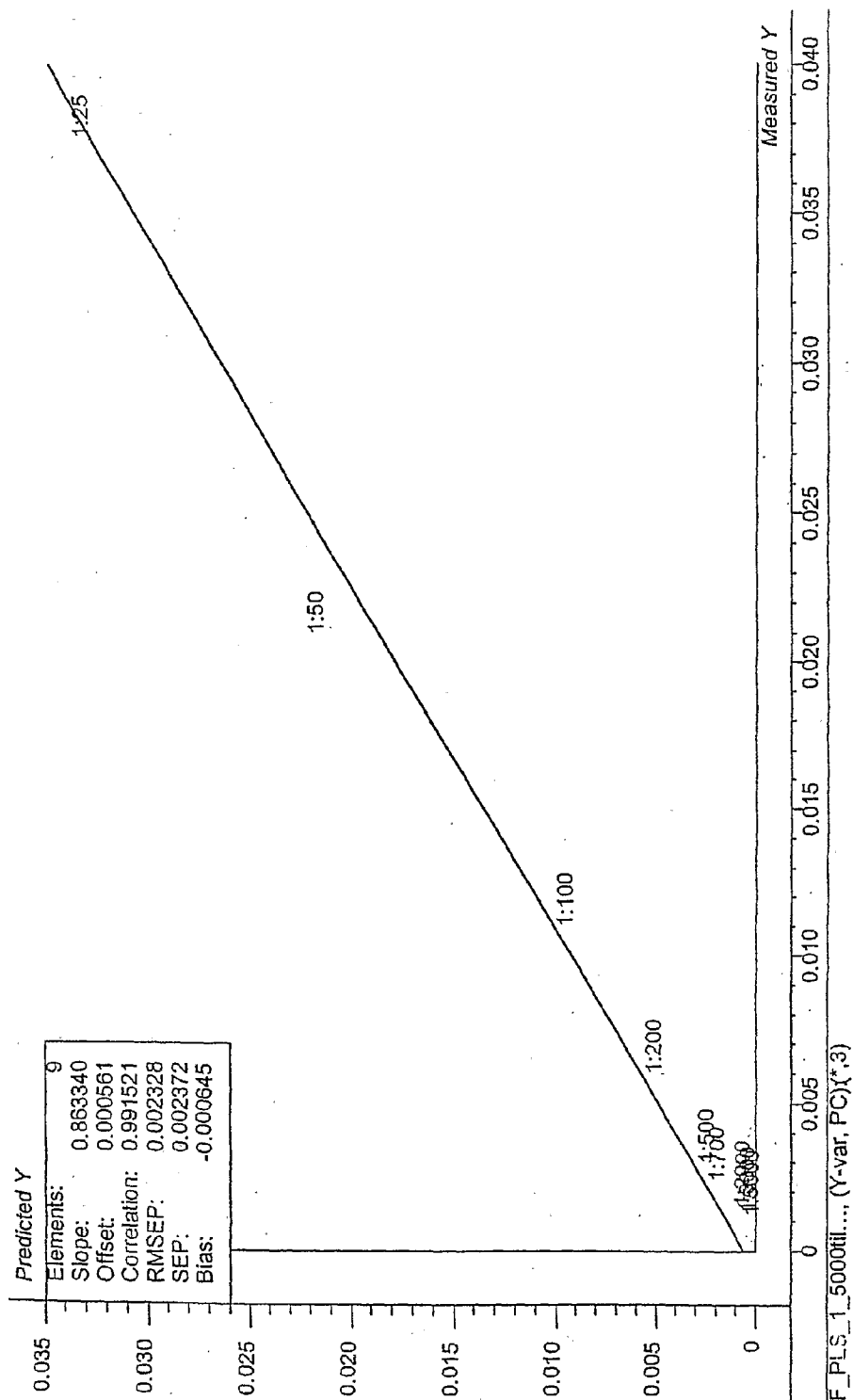


Fig. 21B

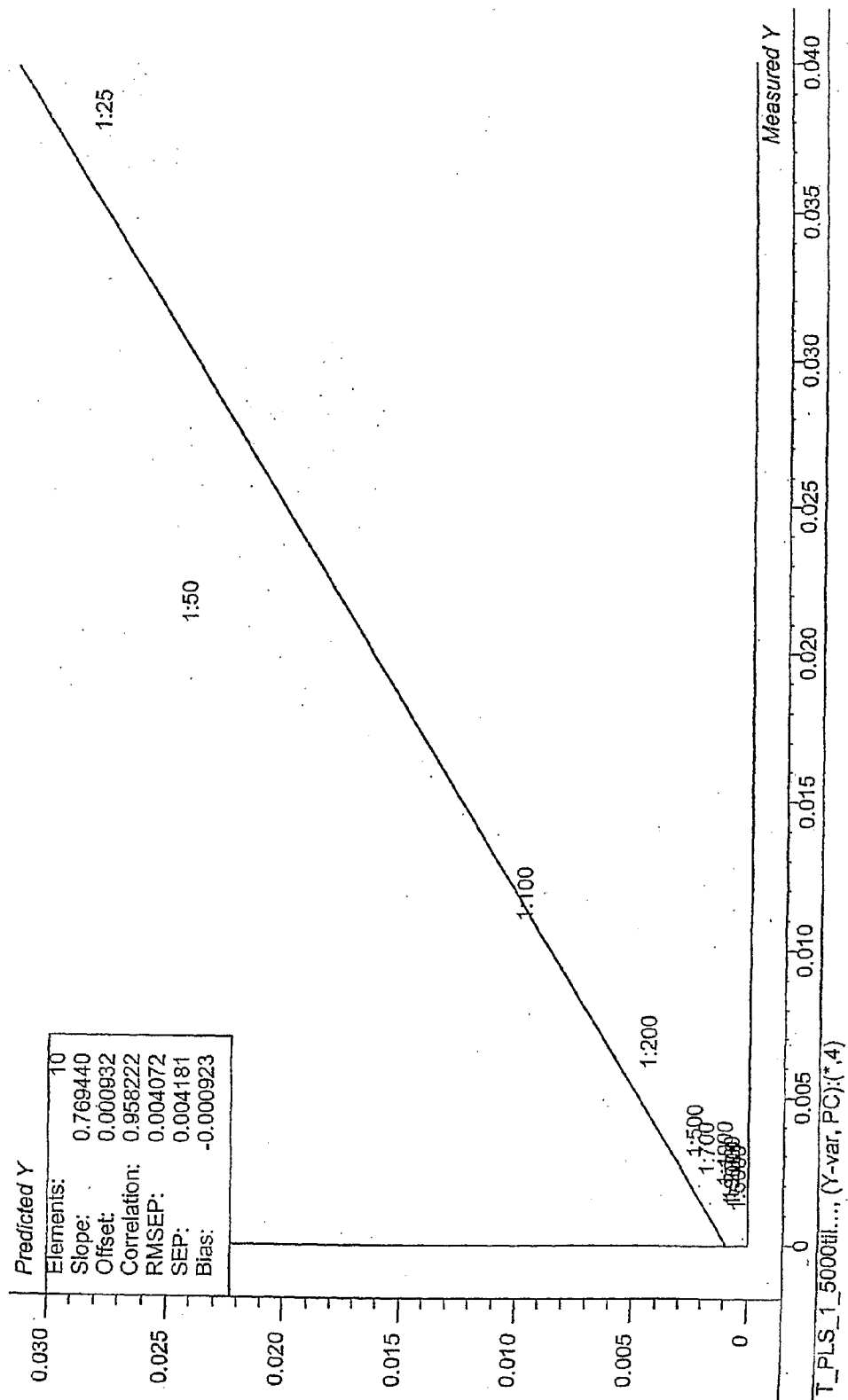


Fig. 22A

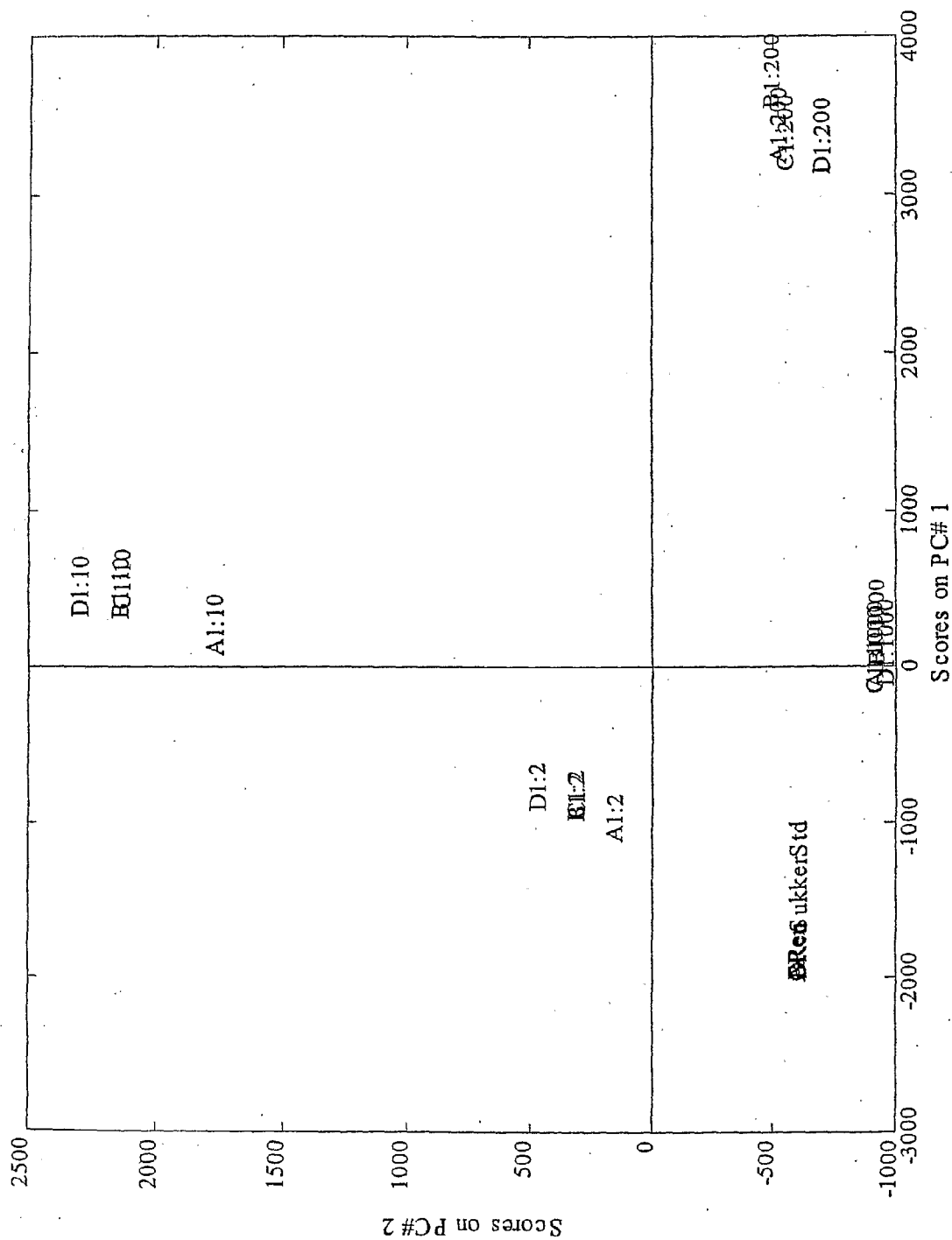
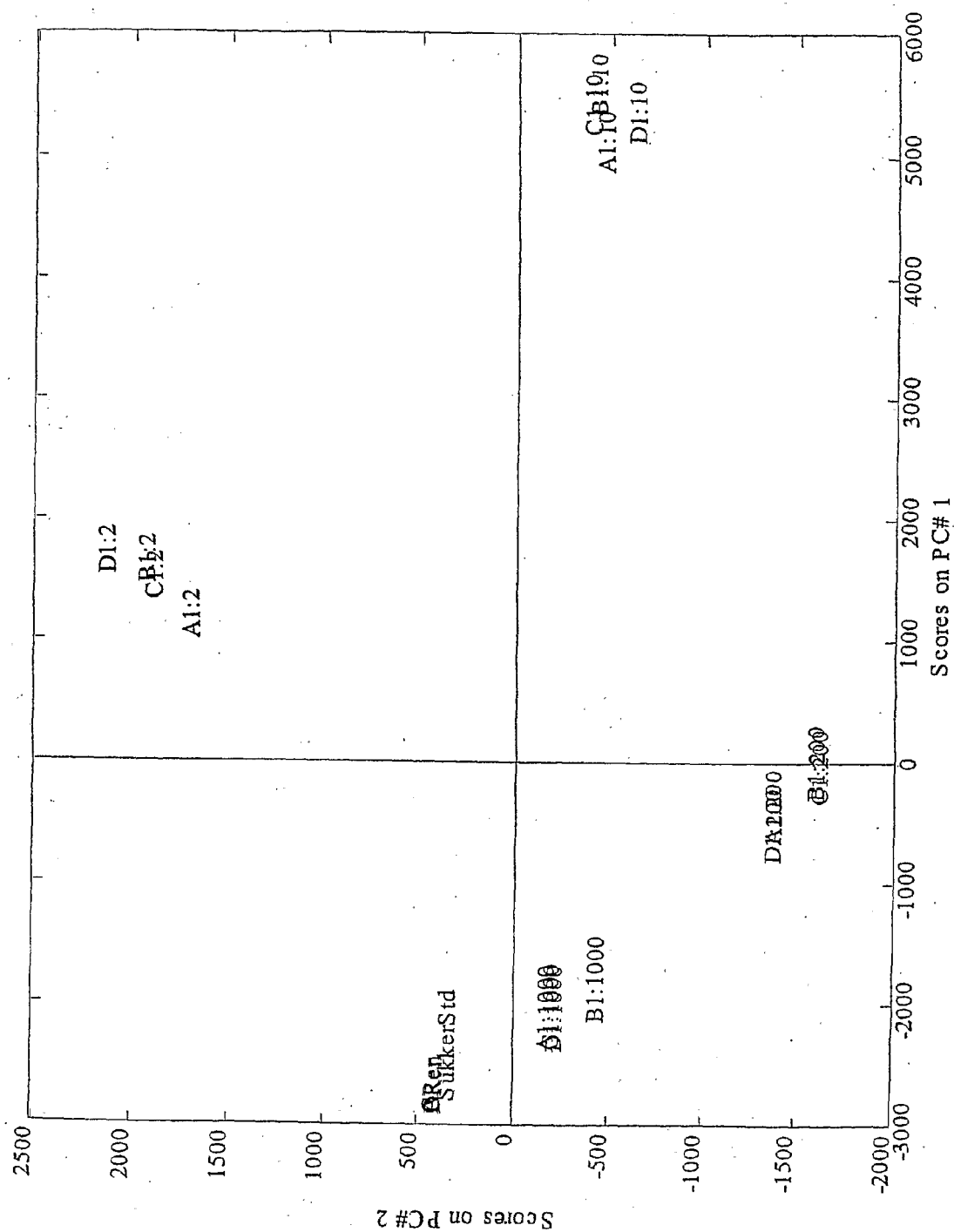
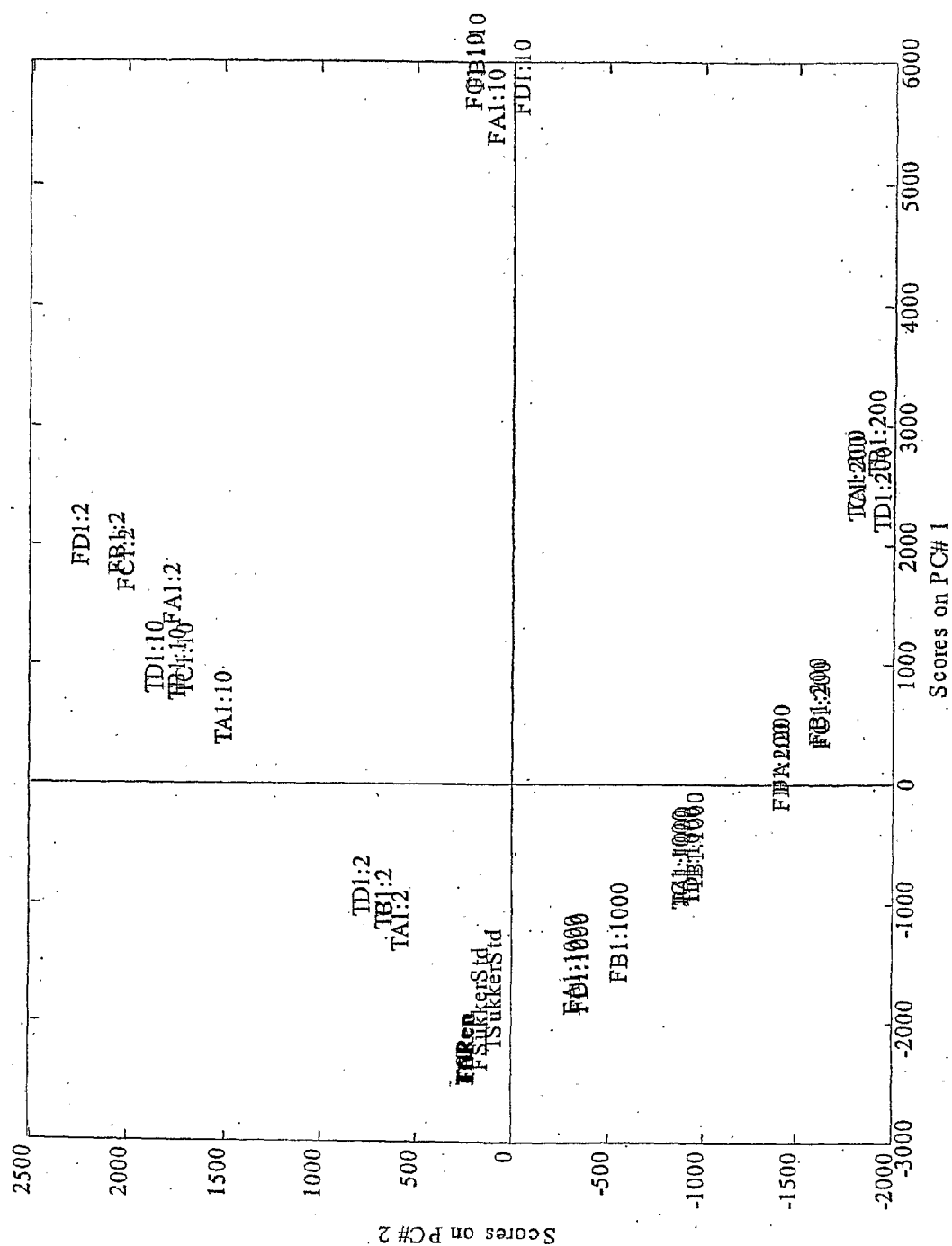


Fig. 22B



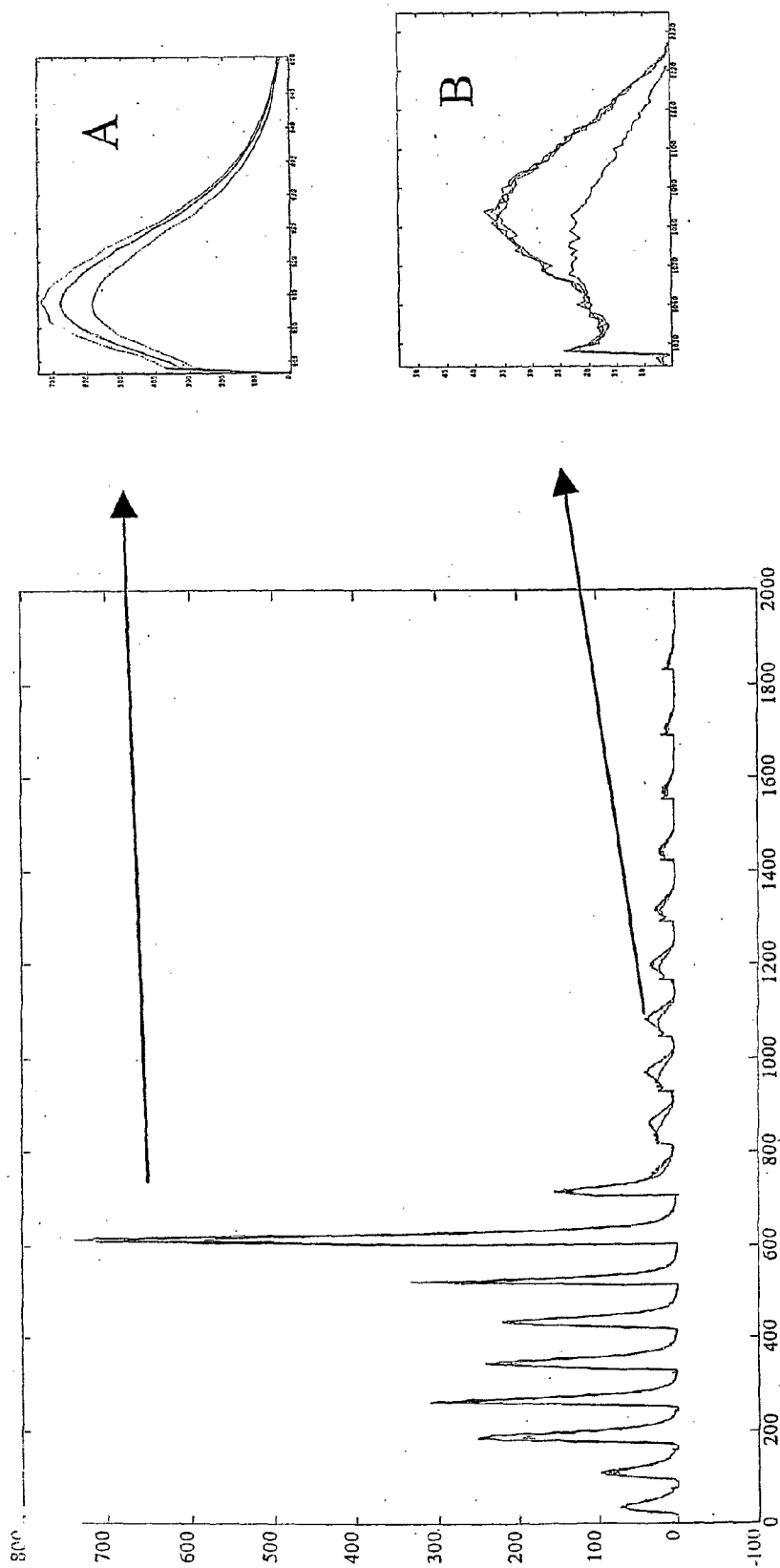
35/36

Fig. 23



36/36

Fig. 24



INTERNATIONAL SEARCH REPORT

Int ☐ onal Application No
PCT/DK 01/00383

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G01N21/64 A61B5/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 7 G01N A61B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, PAJ, WPI Data, INSPEC, COMPENDEX, IBM-TDB

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	<p>US 5 576 544 A (ROSENTHAL ROBERT D) 19 November 1996 (1996-11-19)</p> <p>column 1, line 30 - line 37 column 3, line 52 - column 4, line 62; figure 1</p> <p style="text-align: center;">--- -/-</p>	<p>1-8, 10-29, 31-48, 50-59</p>

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents:

A document defining the general state of the art which is not considered to be of particular relevance

E earlier document but published on or after the international filing date

L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

O document referring to an oral disclosure, use, exhibition or other means

P document published prior to the international filing date but later than the priority date claimed

T later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

X document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

Y document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

8 document member of the same patent family

Date of the actual completion of the international search

4 September 2001

Date of mailing of the international search report

17/09/2001

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040. Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Tabellion, M

INTERNATIONAL SEARCH REPORT

International Application No
PCT/DK 01/00383

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT		
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 96 30746 A (UNIVERSITY OF TEXAS) 3 October 1996 (1996-10-03)	1,2,4, 7-10,12, 13,18, 22,23, 25, 28-30, 32,37, 41,42, 44, 47-51,56
A	abstract page 6, line 28 - line 32 page 7, line 4 - line 5 page 8, line 1 - line 18 page 8, line 26 -page 9, last line page 10, line 15 -page 13, line 14 page 19, paragraph 1 - paragraph 2 page 19, line 35 -page 20, line 8 page 21, line 34 -page 22, line 6 claims 1,2,7,8,10,11; figures 1-4	17,36,55
X	WO 98 24369 A (UNIVERSITY OF TEXAS) 11 June 1998 (1998-06-11)	22,25, 28-32, 41,42, 47-49
A	abstract * page 2, last paragraph - page 3, line 13 * page 5, paragraph 1 page 10, line 11 -page 11, line 21 page 28, line 8 - line 12 page 29, line 9 - line 10 page 37, paragraph 1 claims 1,14,30; figures 2,8	1,2,7-9, 12
X	K.M. O'BRIEN ET AL.: "Development and evaluation of spectral classification algorithms for fluorescence guided laser angioplasty" IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING., vol. 36, no. 4, April 1989 (1989-04), pages 424-4430, XP000031707 IEEE INC. NEW YORK. 111964 1/2, US	22,23, 25, 28-32, 39,41, 42,44, 47-50,58
A	ISSN: 0018-9294 page 424, right-hand column, line 17 - line 19 page 425, left-hand column, line 6 - line 30 page 425, right-hand column, last paragraph -page 426, right-hand column, line 54; figure 1	1,2, 7-10,20

	--- -/--	

INTERNATIONAL SEARCH REPORT

Int. Patent Application No.
PCT/DK 01/00383

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT		
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
E	<p>WO 00 42907 A (INSTRUMENTATION METRICS INC) 27 July 2000 (2000-07-27)</p> <p>page 7, line 22 -page 21, line 16; claims 1-3</p> <p>-----</p>	<p>1-8, 10-29, 31-48, 50-59</p>

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/DK 01/00383

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5576544 A	19-11-1996	US 5204532 A	20-04-1993
		US 5068536 A	26-11-1991
		US 5077476 A	31-12-1991
		US 5086229 A	04-02-1992
		US 5028787 A	02-07-1991
		US 6066847 A	23-05-2000
		AU 2251292 A	12-01-1993
		CA 2111868 A	23-12-1992
		EP 0590077 A	06-04-1994
		JP 6508440 T	22-09-1994
		MX 9202953 A	01-02-1993
		WO 9222804 A	23-12-1992
		AT 145988 T	15-12-1996
		AU 8238791 A	23-01-1992
		CA 2086019 A,C	28-12-1991
		DE 69123448 D	16-01-1997
		DE 69123448 T	22-05-1997
		EP 0536304 A	14-04-1993
		IE 912231 A	01-01-1992
		JP 5508336 T	25-11-1993
		NZ 238717 A	26-08-1994
		WO 9200513 A	09-01-1992
		US 5362966 A	08-11-1994
		US 5436455 A	25-07-1995
		US 5438201 A	01-08-1995
		US 5574283 A	12-11-1996
		US 5237178 A	17-08-1993
		US 5365066 A	15-11-1994
		US 5218207 A	08-06-1993
		ZA 9104977 A	29-04-1992
		AT 163844 T	15-03-1998
		AU 5042190 A	13-08-1990
		CA 2045599 A,C	20-07-1990
		DE 69032126 D	16-04-1998
		DE 69032126 T	15-10-1998
		EP 0456716 A	21-11-1991
		ES 2114860 T	16-06-1998
		JP 2191434 A	27-07-1990
		JP 5058735 B	27-08-1993
		KR 145695 B	01-08-1998
		WO 9007905 A	26-07-1990
WO 9630746 A	03-10-1996	US 5612540 A	18-03-1997
		CA 2215608 A	03-10-1996
		EP 0817959 A	14-01-1998
		JP 11503233 T	23-03-1999
WO 9824369 A	11-06-1998	US 6135965 A	24-10-2000
		EP 0967918 A	05-01-2000
		JP 2001505113 T	17-04-2001
WO 0042907 A	27-07-2000	AU 2620900 A	07-08-2000